UCSF
UC San Francisco Electronic Theses and Dissertations

Title

Transcriptional regulator coding-sequence evolution preceded cis-regulatory changes in the origin of a new transcriptional circuit

Permalink

https://escholarship.org/uc/item/688434g3

Author

Britton, Candace Sarah

Publication Date

2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Transcriptional regulator coding-sequence evolution preceded cis-regulatory changes in the origin of a new transcriptional circuit

by

Candace Britton

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:
Alexander Johnson
C5A8C1DAACAB4A8...

Alexander Johnson

Chair

DocuSigned by:
Geeta Narlikar
DocuSigned by:
043B...

Geeta Narlikar

DocuSigned by:
Hao Li
C2CA172BE8684DD...

Hao Li

Committee Members

Copyright 2019

by

Candace S. Britton

Acknowledgements

I'd like to thank my mentors, colleagues, friends, and family, all of whom contributed to this dissertation and helped keep my spirits up during my PhD. First and foremost, I'm grateful to my advisor, Sandy Johnson. Sandy is a brilliant scientist, and I've learned from his example to formulate ideas precisely, think about data rigorously, and write clearly. Under his mentorship I was able to develop my research independently, with the confidence that he would keep me on a path toward science worth doing. I'm proud of the work that we developed together. My other thesis committee members, Geeta Narlikar and Hao Li, were valuable sources of thoughtful, incisive, and creative questions. They nudged me to clearly conceptualize, explain, and complete my work. Geeta served on every committee and oral exam I had in graduate school, and I'm grateful for her encouragement and coaching along the way.

The Johnson Lab is full of talented and kind scientists, and I'd like to thank them all for their scientific advice and friendship. I've been especially fortunate to have Liron Noiman as my contemporary in the lab. Our desks and lab benches are back-to-back, and we've been at each other's side to interpret every result and process every up and down of graduate school. I'm very grateful for her friendship. I was also fortunate to work with Trevor Sorrells, who mentored me as a rotation student in the lab. Trevor combines an optimistic spirit with strategic thinking in the lab, and the Matα2 evolution project would not have continued without him realizing it would be possible and even had the potential to be surprising. Isabel Nocedal was also a senior graduate student when I joined the laboratory, and her ability to complete a difficult project and make sense of complex experiments was inspiring and instructive. In my opinion, the Johnson Lab has been spoiled with an outstanding string of graduate students, and it's been a pleasure to watch Kyle Fowler take on his research projects with facility, and Francesca del Frate begin conceptualizing a new project in the cell-type specific gene system. I've learned something from all of the

postdoctoral researchers and senior scientists who I have overlapped with in the laboratory as well, including Chiraj Dalal, Matt Lohse, Sheena Singh-Babak, Naomi Ziv, Chien-Der Lee, and Carrie Graham. Matt Lohse has proven to be a source of wisdom that is either completely practical or totally nonsensical. I'm grateful to Chiraj for keeping things in perspective at all times, and for always helping me see how to move forward in research projects and in life. My classmates and other friends in the Tetrad Graduate Program are brilliant, lovable eccentrics, and I've enjoyed all of our conversations, class dinners, text message threads, trips to Giordano Bros., swim workouts, and book club meetings over the past six (plus!) years. I know they'll thrive in their careers beyond UCSF, and I can't wait to see where they end up.

Big news: it turns out graduate school is challenging. I never would have made it through without the support and love of my friends and family. It is no surprise that I wound up in a PhD program given the examples I have in my parents, Carrie and Steven Britton. From my mother I learned the value of asking questions and searching for work that matters. My father is a master hustler who knows the value to be found in effort itself, and I would have gotten even less done in graduate school if it weren't for the good habits I learned from him. My brother Miles is my oldest and dearest friend. Over the last few years he's emerged as an expert alpine mountain guide, and it has been rewarding to see him go through his own intensive educational experience. My grandfather Craig Evans, who is a voracious reader and one of the most data-oriented members of my family, has been especially supportive of my education. I'm grateful for his encouragement and curiosity about my research and science in general. Finally, and centrally, I'm thankful for my boyfriend Seth Adams, because the joy and silliness that we share make everything better.

When I was a high school senior applying to colleges, I was thrilled to learn that you can actually *study* evolutionary biology at a higher level—that there are entire majors and departments dedicated to understanding the deep ancestry that all life on Earth shares, and the ways in which

organisms diversify over time. During my undergraduate education I grew to also appreciate the value of mechanistic biochemical and genetic experiments. I owe a lot to my undergraduate research mentor, Professor Nicole King, for taking me into her lab and showing me how to seek big, new scientific questions. I have been fortunate to be able to bring my old interests together and combine molecular approaches with evolutionary questions in my PhD research. While I have chosen to leave basic research for the next phase of my career, it has been a privilege to practice the rigorous analytical thinking, creative problem solving, biological know-how, and bravery to try new things that are necessary to complete a PhD. I am grateful to the University of California and our taxpayer-funded investments in scientific research and education for giving the opportunity to do so.

The work described in Chapter 2 has been posted as a bioRxiv preprint (https://doi.org/10.1101/565234) and submitted for publication in a peer reviewed journal. It began as a collaboration between Trevor Sorrells and me. Trevor performed preliminary bioinformatic analyses, sequenced the *W. anomalus* α mating-type locus, and conceived of initial experiments to characterize *W. anomalus* Matα2 and the **a**-specific gene *cis*-regulatory elements. I designed all other experiments, performed all of the experiments, and wrote the manuscript in collaboration with Alexander Johnson, who oversaw the project.

Abstract

Transcriptional regulator coding-sequence evolution preceded *cis*-regulatory changes in the origin of a new transcriptional circuit

Candace Britton

Life often diversifies through changes in gene expression patterns. These patterns evolve via changes in transcriptional regulatory circuits that are determined by transcriptional regulatory proteins and the *cis*-regulatory sequences they bind in the genome. While it has long been known that changes in *cis*-regulatory sequences can affect the evolution of gene expression patterns and that transcriptional regulatory proteins can themselves evolve, we know little of how these two types of regulatory changes occur together to generate new circuits. I discerned a stepwise order of evolutionary events in which both regulator protein-coding and *cis*-regulatory changes were necessary to evolve a new transcriptional regulatory circuit (repression of the **a**-specific genes by Matα2 in yeast). The two changes evolved at separate points in time, millions of years apart. First to evolve were coding-sequence changes in the regulator that formed new protein-protein interaction regions. In one lineage, these new protein-protein interactions became necessary for Matα2's ancestral gene regulatory function (repression of the haploid-specific genes with Mat**a**1). In another lineage, millions of years after the coding-sequence changes to Matα2, *cis*-regulatory changes occurred in the **a**-specific genes, thereby co-opting Matα2 for regulation of this new set of target genes. We propose that this evolutionary trajectory is an example of constructive neutral evolution in that Matα2's new protein-protein interactions initially had no consequence to the logic of cell-type specific gene regulation, but eventually allowed for the creation of a novel circuit (Chapter 2). In the course of these investigations, I also observed additional coding-sequence changes in the DNA-binding domain of Matα2 (Chapter 3), and evolutionary changes in the

identities of some of the yeast cell-type specific genes (Chapter 4). The results presented here add to our understanding of the ways in which transcriptional regulatory circuits diversify.

Table of Contents

Chapter 1 Introduction.....1

References 9

Chapter 2 New transcriptional circuit evolved by coding sequence changes in a master regulator followed by *cis*-regulatory changes in its target genes.....12

References.....57

Chapter 3 Multiple regions of the *Wickerhamomyces anomalus* Matα2 DNA-binding domain contribute to its inability to repress **a**-specific gene transcription.....61

References.....76

Chapter 4 *Wickerhamomyces anomalus* cell-type specific genes and their regulation...77

References.....105

Appendix Methods for genetically modifying *Wickerhamomyces anomalus* and other Phaffomycetaceae species 107

References.....113

List of Tables

Chapter 3

Table 1. *S. cerevisiae* strains used in this chapter.....73

Chapter 4

Table 1. Complete list of Mat**a**1-Matα2 regulated genes in *W. anomalus* **a**/α cell ...98

Table 2. Regions reproducibly enriched in HA-Mat**a**1 and HA-Matα2 chromatin immunoprecipitation.....99

Table 3. *W. anomalus* strains used in this chapter.....100

Appendix

Table 1. Recommended concentrations of selective drugs for transforming various Phaffomycetaceae species and strains.....112

List of Figures

Chapter 2

Figure 1. Cell-type specific gene expression in the Saccharomycotina yeast 24

Figure 2. *W. anomalus* Matα2 has functional Tup1 and Mcm1-interacting regions but does not repress the **a**-specific genes 25

Figure 3. Mat**a**1, Matα2, and a Mcm1 cis-regulatory sequence are all required for haploid-specific gene repression in *W. anomalus* 27

Figure 4. Order of evolutionary events leading to repression of the **a**-specific genes by Matα2, mapped on to the yeast cladogram 29

Supp. Figure 1. *W. anomalus* clade Matα2 orthologs contain functional Tup1 and Mcm1-interaction regions, but cannot repress **a**-specific gene transcription due to changes in homeodomain.....35

Supp. Figure 2. Cell-type specific gene expression patterns in *W. anomalus* 37

Supp. Figure 3. *W. anomalus MATα2* does not directly repress the **a**-specific genes 38

Supp. Figure 4. *W. anomalus* **a**-specific genes are activated by Mat**a**2, the ancestral mode of gene regulation 40

Supp. Figure 5. The *W. anomalus* haploid-specific genes are repressed by Matα2 and Mat**a**1.....42

Supp. Figure 6. Mcm1, Matα2, and Mat**a**1 are directly required for *RME1* repression in **a**/α cell.....44

Chapter 3

Figure 1. Only *W. anomalus*-clade Matα2 has DNA-binding domains that does

not support **a**-specific gene repression....70

Figure 2. Alignment of *S. cerevisiae* and *W. anomalus* Matα2 homeodomains....71

Figure 3. N-terminal arm is greater contributor to *W. anomalus* Matα2 DNA-binding domain's inability to repress **a**-specific gene transcription....72

Chapter 4

Figure 1. Cell-type specific gene regulation in *W. anomalus*....89

Figure 2. Relative expression of Matα2 and Mat**a**1-regulated genes across wildtype *W. anomalus* cell types....90

Figure 3. Expression of categorized Matα2 and Mat**a**1-regulated genes in **a**/α cells with *MATα2* or *MAT**a**1* deleted....91

Figure 4. *W. anomalus* α-specific gene *cis*-regulatory sequences....93

Figure 5. *W. anomalus* **a**-specific gene *cis*-regulatory sequences....95

Figure 6. *AXL1* motif scoring....97

Chapter One

Introduction

Changes in gene expression patterns are an integral way in which organisms evolve. Biologists have found time and time again that what makes two related organisms different from each other is not simply what genes they have or the sequences of those genes, but instead how those genes are actually deployed in gene expression patterns. Emile Zuckerkandl and Linus Pauling were among the first to propose that the timing and synthesis rates of so-called "informational macromolecules" would affect phenotypes, in addition to actual sequences of structural genes (Zuckerkandl and Pauling, 1965). Their insight proved to be prescient, and the importance of regulatory evolution was apparent from early comparisons of protein sequences from related species. In particular, human and chimpanzee protein sequences turned out to be so similar that Mary-Claire King and Allan Wilson immediately interpreted their great similarity as evidence for Susumo Ohno's proposal that regulatory evolution, rather than coding-sequence changes to proteins, accounts for most morphological differences in organisms (King and Wilson, 1975; Ohno, 1972). Evolution through changes in regulatory systems, Ohno wrote, can explain the disconnect between the great diversity of the morphology of organisms on Earth, and the apparent difficulty with which individual macromolecules tolerate mutations (Ohno, 1972).

Since then, the importance of regulatory evolution in eukaryotes has been born out in countless examples, both large and small. We now know that regulatory evolution primarily occurs at the level of transcriptional regulation—that is, changes in the relative activation or repression of the transcription of specific mRNAs (Carroll, 2005; 2008; Erwin and Davidson, 2009). This process is governed by sequence-specific DNA-binding proteins (or, "transcriptional regulators") that bind DNA at *cis*-regulatory sequences in the genome and effect the transcription of target genes (Alberts et al., 2014). The features of transcriptional regulation (i.e., *cis*-regulatory sequences and sequence-specific DNA-binding proteins) are the grist for the mill of regulatory evolution. For example, there can be phenotypic consequences from even simple changes such as

the gain or loss of a single *cis*-regulatory site that brings a gene in or out of control of a certain transcriptional regulator protein (Jiang and Rausher, 2018; Prud'homme et al., 2006; Wang et al., 1995). At the other extreme, transcriptional circuits can undergo sweeping rewiring events, with regulator proteins gaining and losing control hundreds of target genes, sometimes evolving to regulate entirely different processes in related organisms (Nocedal et al., 2017; Schmidt et al., 2010; Stefflova et al., 2013; Tuch et al., 2008). Both of these extreme forms of regulatory evolution are explained by the gain or loss of *cis*-regulatory sequences, whether one or thousands.

Indeed, an opinion quickly took hold in the maturation of the gene regulatory evolution field: that *cis*-regulatory mutations are the most important and/or prevalent sources of such gene expression changes, rather than regulatory proteins themselves (King and Wilson, 1975). The rationale was that because mutations in a regulator protein could alter (and most likely, disrupt) its ability to regulate transcription, which could in turn have off-target consequences on the regulation of many genes at once. The chances of pleiotropic, deleterious effects of mutating a regulator protein pointed to *cis*-regulatory mutations as a more favorable driver of gene regulatory evolution. Furthermore, because *cis*-regulatory sequences are both shorter in sequence and more restricted in their functional roles in an organism, their mutations would be more likely to occur and could be more easily tolerated (Carroll, 2005; 2008). This view became increasingly codified, with multiple scholars making the case for the importance and prevalence of *cis*-regulatory mutations (Stern and Orgogozo, 2008; Wittkopp and Kalay, 2011; Wray, 2007). There are many examples of *cis*-regulatory mutations mattering for the evolution of traits, and there is no denying their importance for regulatory evolution.

The importance of *cis*-regulatory changes, however, does not exclude the importance of regulator proteins themselves. That is, the fact that *cis*-regulatory changes are prevalent and important does not mean that protein-coding changes in transcriptional regulators are not also

important (Cheatle Jarvela and Hinman, 2015; Hoekstra and Coyne, 2007; Hsia and McGinnis, 2003; Lynch and Wagner, 2008). There are also many examples of the coding-sequences of transcriptional regulatory proteins evolving. For one, while the DNA-binding specificities of many transcriptional regulatory proteins is deeply conserved, the specificities of some sequence-specific DNA-binding proteins change over time and can co-evolve with their binding sites (Baker et al., 2011; Sayou et al., 2014; Wynne and Treisman, 1992). In addition, transcriptional regulatory proteins can gain and lose amino acid regions that interact with other proteins or are post-translationally modified (e.g., by phosphorylation) (Baker et al., 2012; Lynch et al., 2011; Nnamani et al., 2016). Coding-sequence changes in transcriptional regulators can be associated with new transcriptional circuits, and new organismal phenotypes (Baker et al., 2012; Galant and Carroll, 2002; Ronshaugen et al., 2002). Thus, *cis*-regulatory mutations may indeed be a more prevalent way for transcriptional circuits to evolve, especially over relatively short evolutionary timespans. This does not mean, though, that coding-sequence changes to transcriptional regulatory proteins are *unimportant.* Indeed, the fact that these evolutionary changes do happen and have consequences for gene expression regulation is sufficient for them to be important for evolution. Their importance have been underestimated because these protein-coding changes evolve over the course of tens or hundreds of millions of years, rather than the less than ten million years that separates closely related organisms like humans and chimpanzees. If we accept the importance of coding-sequence changes in transcriptional regulators, then new sets of specific questions about their evolution emerges. How do regulatory proteins evolve in ways that do not have deleterious consequences for their gene targets? It is already known that evolutionary changes to transcriptional regulator proteins are often possible because transcriptional regulatory proteins are modular in structure (Cheatle Jarvela and Hinman, 2015; Lynch and Wagner, 2008). That is, changes in one region of a protein need not disrupt its fold and may not affect its DNA-binding at

all. In addition, given that both *cis*-regulatory and regulator protein-coding changes happen, how do the two evolve in concert? Transcriptional regulatory proteins, like all proteins, evolve via pathways that are constrained by their functions: not all evolutionary trajectories are equally possible, because some paths through sequence space lead to non-functional, deleterious intermediates (Harms and Thornton, 2013; Weinreich et al., 2006). How, then, do the principles of protein evolution shape the actual evolutionary trajectory of a new transcriptional circuit that required both regulator protein-coding and non-coding *cis*-regulatory changes?

The cell-type specific genes in the Saccharomycotina fungi have proven to be a fruitful model for studying the evolution of transcriptional regulation (Baker et al., 2012; Booth et al., 2011; Sorrells et al., 2015; Tsong et al., 2003; 2006). These gene expression patterns are small sets of genes specific to the separate mating cell types in Saccharomycotina yeast (i.e., the two haploid mating types compared to the diploid non-mating cell type), and their proper expression is essential for successful mating (Herskowitz, 1989). The cell-type specific gene expression patterns are governed by four transcriptional regulatory proteins encoded by the mating-type locus. These transcriptional regulatory proteins have been studied for decades as pioneering models for eukaryotic transcriptional regulation itself. The depth of the previous research into their mechanisms of transcriptional regulation and their suitability as model systems transformed them into excellent models for expanding our understanding of how transcriptional regulation evolves. For example, the mating-type transcriptional regulators are non-essential genes, so they are relatively easy to genetically manipulate. Several of the mating type transcriptional regulator proteins and their co-regulatory proteins have been purified and have solved X-ray crystallography structures. Their DNA-binding specificity is known, as are their binding sites are in the genome, which genes they regulate, which of their amino acid residues are necessary for binding DNA or co-regulatory protein, and what their tolerance is for certain amino acid substitutions (Galgoczy et

al., 2004). This rich literature allows *S. cerevisiae* to be a launching pad to study different branches of the yeast family tree. By comparing the mating-type transcriptional regulators and the transcriptional circuits that they are members of to their homologous proteins and circuits in *S. cerevisiae*, we can ask specific, mechanistic questions about how their evolution (Dalal and Johnson, 2017; Nocedal and Johnson, 2015; Sorrells and Johnson, 2015). The Saccharomycotina yeast represent approximately 300-500 million years of evolution (roughly equivalent to the divergence time between most animals), so this approach allows us to understand ancient evolutionary transitions in a deeply mechanistic way (Shen et al., 2018).

In this thesis, I ask how regulator protein-coding and *cis*-regulatory evolutionary changes evolve together to produce a new gene regulatory circuit. My model was the transcriptional regulatory protein Matα2, which gained control of one of the yeast cell-type specific gene expression patterns (the **a**-specific genes) over the course of hundreds of millions of years. As presented in Chapter 2, I discovered that direct **a**-specific gene regulation by Matα2 evolved in a stepwise fashion. First, the Matα2 protein-coding sequence changed, gaining two stretches of amino acid residues that are capable of forming specific protein-protein contacts with other transcriptional regulatory proteins. In one lineage (the branch of the yeast tree leading to modern-day *Wickerhamomyces anomalus*), these new interactions functioned in Matα2's ancient, conserved gene expression function (repression of the haploid-specific genes in **a**/α cells with Mat**a**1). Millions of years later, these new interactions were co-opted for the direct regulation of the **a**-specific genes. This second change occurred in the lineage leading to *S. cerevisiae* and required the *cis*-regulatory gain of Matα2-Mcm1 binding sites at the **a**-specific genes. Thus, a new transcriptional circuit evolved first via the gain of functionally-neural coding-changes to a regulator protein, followed by *cis*-regulatory changes that recruited it to new target genes.

To answer these questions, I studied Matα2's function in the non-model yeast species *Wickerhamomyces anomalus*. This species branches outside of *S. cerevisiae* and its closer relatives in the Saccharomycetaceae clade, making it phylogenetically relevant for parsing Matα2 and **a**-specific gene regulatory evolution (Shen et al., 2018; 2016). In the course of these investigations, I also discovered coding-sequence changes to the *W. anomalus* Matα2 DNA-binding domain that make it not capable of **a**-specific gene repression. These results are presented in Chapter 3. I further sought to fully catalog the cell-type specific genes in *W. anomalus* and found suggestions of a new possible transcriptional circuit re-wiring in haploid-specific gene regulation in *W. anomalus*. These results are presented in Chapter 4. Finally, I found a protocol suitable for genetically altering *W. anomalus* and developed a second drug resistance marker for selection *W. anomalus*, as presented in the Appendix (Schorsch et al., 2009).

Studying the evolution of yeast cell-type specific gene regulation has proven the power of questioning assumptions about evolution. By pursuing an understanding of how one transcriptional regulator protein evolved, we have found that the old objections to transcriptional regulator proteins as an important and prevalent source of evolutionary change (i.e., that they are too complex and may have too many off-target effects to readily evolve) may indeed be real constraints. Nonetheless, Matα2 has diversified and gained new regulatory functions *within* those constraints. Indeed, the constraints of maintaining its ancestral function seem to have been essential to its evolutionary gains in the first place. Evolving new regions of the protein that form specific protein-protein interactions, but only using those interactions at first to complete Matα2's ancestral function, created the opportunity for a new function to later emerge. Matα2 is a homeodomain protein, a deeply conserved group of transcriptional regulators in eukaryotes thought to have been present in the last eukaryotic common ancestor (Derelle et al., 2007). Homeodomain proteins tend to regulate gene expression patterns involved in morphological

processes, especially in animals, where the homeotic Hox genes are homeodomain transcription factors (2011). My findings about Matα2's evolution may prove to be the first of many examples of homeodomain proteins evolving new functions without disrupting more ancient ones.

References

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2014). Molecular Biology of the Cell (Garland Publishing).

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

Baker, C.R., Tuch, B.B., and Johnson, A.D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc. Natl. Acad. Sci. U.S.a. *108*, 7493–7498.

Booth, L.N., Tuch, B.B., and Johnson, A.D. (2011). Intercalation of a new tier of transcription regulation into an ancient circuit. Nature *468*, 959–963.

Carroll, S.B. (2005). Evolution at two levels: on genes and form. PLoS Biol *3*, e245.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36.

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

Dalal, C.K., and Johnson, A.D. (2017). How transcription circuits explore alternative architectures while maintaining overall circuit output. Genes Dev. *31*, 1397–1405.

Derelle, R., Lopez, P., Le Guyader, H., and Manuel, M. (2007). Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. Evol. Dev. *9*, 212–219.

Erwin, D.H., and Davidson, E.H. (2009). The evolution of hierarchical gene regulatory networks. Nat Rev Genet *10*, 141–148.

Galant, R., and Carroll, S.B. (2002). Evolution of a transcriptional repression domain in an insect Hox protein. Nature *415*, 910–913.

Galgoczy, D.J., Cassidy-Stone, A., Llinás, M., O'Rourke, S.M., Herskowitz, I., DeRisi, J.L., and Johnson, A.D. (2004). Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.a. *101*, 18069–18074.

Harms, M.J., and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet *14*, 559–571.

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

Hoekstra, H.E., and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. Evolution *61*, 995–1016.

Hsia, C.C., and McGinnis, W. (2003). Evolution of transcription factor function. Curr. Opin. Genet. Dev. *13*, 199–206.

Jiang, P., and Rausher, M. (2018). Two genetic changes in cis-regulatory elements caused evolution of petal spot position in Clarkia. Nat Plants *4*, 14–22.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116.

Lynch, V.J., and Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. Evolution *62*, 2131–2154.

Lynch, V.J., May, G., and Wagner, G.P. (2011). Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. Nature *480*, 383–386.

Nnamani, M.C., Ganguly, S., Erkenbrack, E.M., Lynch, V.J., Mizoue, L.S., Tong, Y., Darling, H.L., Fuxreiter, M., Meiler, J., and Wagner, G.P. (2016). A Derived Allosteric Switch Underlies the Evolution of Conditional Cooperativity between HOXA11 and FOXO1. Cell Rep *15*, 2097–2108.

Nocedal, I., and Johnson, A.D. (2015). How Transcription Networks Evolve and Produce Biological Novelty. Cold Spring Harbor Symposia on Quantitative Biology *80*, 265–274.

Nocedal, I., Mancera, E., and Johnson, A.D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. Elife *6*, 1.

Ohno, S. (1972). An argument for the genetic simplicity of man and other mammals. Journal of Human Evolution *1*, 651–662.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M., Yeh, S.-D., True, J.R., and Carroll, S.B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature *440*, 1050–1053.

Ronshaugen, M., McGinnis, N., and McGinnis, W. (2002). Hox protein mutation and macroevolution of the insect body plan. Nature *415*, 914–917.

Sayou, C., Monniaux, M., Nanao, M.H., and Moyroud, E. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science *328*, 1036–1040.

Schorsch, C., Köhler, T., and Boles, E. (2009). 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.

Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell *175*, 1533–1545.e20.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C.P., Hittinger, C.T., and Rokas, A. (2016). Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. G3 (Bethesda) *6*, 3927–3939.

Sorrells, T.R., and Johnson, A.D. (2015). Making sense of transcription networks. Cell *161*, 714–723.

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

Stefflova, K., Thybert, D., Wilson, M.D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D.J., Talianidis, I., Marioni, J.C., et al. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell *154*, 530–540.

Stern, D.L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? Evolution *62*, 2155–2177.

Tsong, A.E., Miller, M.G., Raisner, R.M., and Johnson, A.D. (2003). Evolution of a Combinatorial Transcriptional Circuit. Cell *115*, 389–399.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative transcriptional circuits with identical logic. Nature *443*, 415–420.

Tuch, B.B., Galgoczy, D.J., Hernday, A.D., Li, H., and Johnson, A.D. (2008). The evolution of combinatorial gene regulation in fungi. PLoS Biol *6*, e38.

Wang, Y., Harvey, C.B., Pratt, W.S., Sams, V.R., Sarner, M., Rossi, M., Auricchio, S., and Swallow, D.M. (1995). The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. Hum. Mol. Genet. *4*, 657–662.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science *312*, 111–114.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet *13*, 59–69.

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

Wynne, J., and Treisman, R. (1992). SRF and MCM1 have related but distinct DNA binding specificities. Nucleic Acids Res. *20*, 3297–3303.

Zuckerkandl, E., and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. In Evolving Genes and Proteins, (Elsevier), pp. 97–166.

(2011). A Handbook of Transcription Factors (Dordrecht: Springer Netherlands).

Chapter Two

New transcriptional circuit evolved by coding sequence changes in a master regulator followed by *cis*-regulatory changes in its target genes

Abstract

While changes in both the coding-sequence of transcriptional regulators and in the *cis*-regulatory sequences recognized by them have been implicated in the evolution of transcriptional circuits, little is known of how they evolve in concert. We describe an evolutionary pathway in fungi where a new transcriptional circuit (**a**-specific gene repression by Matα2) evolved by coding changes in an ancient master regulator, followed millions of years later by *cis*-regulatory sequence changes in the genes of its future regulon. We discerned this order of events by analyzing a group of species in which the coding changes in the regulator are present, but the *cis*-regulatory changes in the target genes are not. In this group we show that the coding changes became necessary for the regulator's deeply conserved function and were therefore preserved. We propose that the changes first arose without altering the overall function of the regulator (although changing the details of its mechanism) and were later co-opted to "jump start" the formation of the new circuit.

Introduction

Changes in transcriptional circuits over evolutionary time are an important source of organismal novelty. Such circuits are typically composed of one or more transcriptional regulators (sequence-specific DNA-binding proteins) and their direct target genes, which contain *cis*-regulatory sequences recognized by the regulators. Although much early emphasis was placed on changes of *cis*-regulatory sequences as a source of novelty, it is now clear that coding-changes in the transcriptional regulatory proteins that bind the *cis*-regulatory sequences are also of key importance (Cheatle Jarvela and Hinman, 2015; Lynch and Wagner, 2008; Stern and Orgogozo, 2008; Wittkopp and Kalay, 2011; Wray, 2007). Indeed, some well-documented changes in transcriptional circuitry require concerted changes in both *cis*-regulatory sequences and the regulators that bind them (Baker et al., 2011; Sayou et al., 2014). Although such concerted changes are likely to be widespread, we know little about how they occur.

Here, we study a case in the fungal lineage where gains in *cis*-regulatory sequences in a set of target genes as well as coding changes in the master transcriptional regulator itself are needed for a new circuit to evolve. Specifically, we addressed which came first, the changes in the regulatory protein or the changes in the *cis*-regulatory sequences of its 5-10 target genes. We document that the coding-sequence changes in the regulatory protein occurred millions of years before the gains of the *cis*-regulatory sequences. This observation poses an apparent paradox: how could the protein acquire the needed changes when the new circuit was completed only much later in evolutionary time? We answer this question by revealing a new, unanticipated regulatory scheme present in a small clade of understudied fungal species that readily accounts for this evolutionary trajectory.

Results

The expression of the cell-type specific genes in fungi provides a powerful system to mechanistically study the evolution of gene expression patterns across a period of several hundred million years (Booth et al., 2010; Dalal et al., 2016; Lavoie et al., 2010; Sorrells et al., 2015; Tsong et al., 2003a; 2006; Tuch et al., 2008; Baker et al., 2012). The Saccharomycotina clade of fungi, which spans this range of evolutionary time (see Fig. 1), have two types of mating cells, **a** and α, each of which express a set of genes specific to that cell type. Only **a** cells express **a**-specific genes and only α cells express α-specific genes. Both **a** and α cells express an additional set of genes, the haploid-specific genes (Fig. 1a). The proteins encoded by these sets of genes have been intensively studied and are critical for the sexual cycle (mating and meiosis) of these species (see Supplementary Information for more details).

The expression patterns of these three sets of genes are broadly conserved across extant yeasts (as are the genes themselves) and therefore must have been present in the last common ancestor of these species, some hundreds of millions of years ago. However, extant species

accomplish the regulation of the cell-type specific genes using different mechanisms (Fig. 1b) (Tsong et al., 2003a; 2006; Baker et al., 2012). For example, in one group of species, the **a**-specific genes are turned on by a transcriptional activator protein (Mat**a**2) that is made only in **a** cells and which binds a *cis*-regulatory sequence upstream of the **a**-specific genes. In another group of species the **a**-specific genes are turned on by a ubiquitous activator and turned off by binding of a repressor (Matα2) that is made only in α and **a**/α cells. Both schemes ensure that the **a**-specific genes are expressed only in **a** cells, although the mechanism behind this pattern is completely different. Previous work established that the activation scheme (activating **a**-specific genes by Mat**a**2) is ancestral to the repression scheme (repression of the **a**-specific genes by Matα2) (Fig. 1b)(Tsong et al., 2003a).

The switch between these two mechanisms of controlling the **a**-specific genes was completed sometime before the divergence of *Saccharomyces cerevisiae* and *Kluyveromyces lactis* (formally known as Saccharomycetaceae, here called the *S. cerevisiae* clade) but after the divergence of this clade and that containing *Candida albicans* and *Pichia membrifaciens* (formally, Pichiaceae and Debaryomecetaceae, here called the *C. albicans* clade) (Fig. 1b). Three events must have occurred for the newer (repression) scheme to have evolved: (1) Matα2 acquired the ability to contact the Tup1-Ssn6 co-repressor, bringing it to DNA to carry out the repression function, (2) Matα2 acquired the ability to bind to DNA cooperatively (through a direct protein-protein contact) with Mcm1, a ubiquitously expressed transcription regulator necessary for **a**-specific gene repression, and (3) the **a**-specific genes (numbering between 5 and 10, depending on the species) acquired *cis*-regulatory sites for Matα2 (Fig. 1b).

To determine the order of these events, we studied Matα2 and the regulation of the **a**-specific genes in a clade that branched from the ancestor before the completion of all three of these events; that is, before the **a**-specific genes had been brought under negative control by Matα2. We

reasoned that this group of species might have acquired some, but not all of the changes needed to form the new circuit and therefore might provide clues to its evolutionary history. This approach was made possible by the sequencing of genomes of a monophyletic group of species that branches outside of the last common ancestor of the *S. cerevisiae* clade (formally known as the Phaffomycetaceae) (Fig. 1b) (Riley et al., 2016; Shen et al., 2016). We chose the species *Wickerhamomyces anomalus* as a representative of this branch because it does not undergo mating-type switching, greatly simplifying its analysis, and because we worked out relatively simple procedures to genetically alter it (Kurtzman and Fell, 1998).

The central player in the evolutionary pathway investigated here is Matα2, a homeodomain protein. In *S. cerevisiae*, this protein represses both the **a**-specific genes (in combination with Mcm1, a MADS box protein) and the haploid-specific genes (in combination with Mat**a**1, another homeodomain protein). The latter circuit is ancient, present in all three clades, whereas the former, as discussed above, is a relatively recent acquisition, found only in the *S. cerevisiae* clade (Fig. 1b). The *S. cerevisiae* protein has a modular organization of five functional regions (Fig. 2a). Proceeding from N to C-terminus, these are (1) an N-terminal region that binds the transcriptional co-repressor protein Tup1-Ssn6, (2) a domain involved in heterodimerizing with Mat**a**1 to regulate the haploid-specific genes, (3) an approximately 10 amino acid stretch that binds Mcm1, so that Matα2 and Mcm1 bind cooperatively to the *cis*-regulatory sequences controlling the **a**-specific genes, (4) the homeodomain, required to bind *cis*-regulatory sequences in both the **a**-specific and haploid-specific genes, and (5) a C-terminal region that binds Mat**a**1 and is needed for Mat**a**1 and Matα2 to jointly bind DNA (Ho et al., 1994; KOMACHI et al., 1994; Komachi, 1996; Li et al., 1998; Mak and Johnson, 1993; Mead et al., 2002; PHILLIPS et al., 1994; Sauer et al., 1988; Tan and Richmond, 1998; Tsong et al., 2003b). *C. albicans* Matα2, which represents the ancestral state of the Matα2 coding sequence, lacks the region that interacts with Tup1 (region 1) and the residues

that interact with Mcm1 (region 3), but shows a significant similarity to *S. cerevisiae* clade Matα2s in the other three regions (Supplementary Fig. 1) (Baker et al., 2012; Tsong et al., 2003b; 2006). We previously tested and confirmed these predictions experimentally: when the *C. albicans* Matα2 protein is expressed in *S. cerevisiae*, it cannot repress the *S. cerevisiae* **a**-specific genes(Baker et al., 2012). However, if regions 1 and 3 from the *S. cerevisiae* Matα2 are grafted onto the *C. albicans* Matα2, it can now repress these genes (Baker et al., 2012).

We examined the *W. anomalus* Matα2 protein sequence to determine whether it is more similar to the ancestral (represented by *C. albicans)*, or the derived (represented by *S. cerevisiae*) form of Matα2. Alignment of the coding sequence of *W. anomalus* Matα2 with orthologs of Matα2 from many other species indicated that it shares all five functional regions with the *S. cerevisiae* protein (Supplementary Fig. 1). In particular, it has a similar region 1 (the Tup1-interacting region) and a similar region 3 (the Mcm1-interacting region), the regions that the *C. albicans* protein lacks. By swapping these *W. anomalus* regions into the *S. cerevisiae* protein, we confirmed that they are functional in repressing the **a**-specific genes (Fig. 2b). In the course of these experiments, we found, unexpectedly, that the homeodomain of the *W. anomalus* protein contained mutations that prevented its binding to the **a**-specific gene *cis*-regulatory sequence in *S. cerevisiae*, a derived change within this clade alone that we discuss below (Fig. 2b, Supplementary Fig. 1). Similar results were obtained with Matα2 protein from two additional species that branch with *W. anomalus*, indicating that these two conclusions (*W. anomalus* clade Matα2 has functional regions 1 and 3, but cannot bind the *S. cerevisiae* **a**-specific genes) are characteristic of the *W. anomalus* clade rather than a single species (Supplementary Fig. 1d).

The fact that the *W. anomalus* Matα2 protein had acquired the necessary coding changes to interact with Tup1 and Mcm1 but could not bind to the *S. cerevisiae* **a**-specific gene control region raised the question of whether it has a role in regulating the **a**-specific genes in *W. anomalus*.

First, we note that the *W. anomalus* **a**-specific genes are indeed cell-type regulated: they are only expressed in **a** cells (Supplementary Fig. 2). In principle, there are three mechanisms that can produce proper expression of these genes: (1) the **a**-specific genes could be under positive regulation by the transcriptional activator Mat**a**2, as is the situation in *C. albicans* and other outgroup species. Alternatively, (2) the **a**-specific genes could be under negative regulation by Matα2 as they are in *S. cerevisiae*, or (3) the **a**-specific genes could be under both positive regulation by Mat**a**2 and negative regulation by Matα2 as seen in *Lachancea kluyveri* and *Kluyveromyces wickerhamii*(Baker et al., 2012). We determined that the first possibility is correct. Specifically, we tested the effects of deleting Matα2 from an α cell and Mat**a**2 from an **a** cell on cell-type specific gene expression and mating. We found that Mat**a**2 is required for **a**-specific gene activation and for **a** cells to mate, but that Matα2 does not repress the **a**-specific genes and is not required for α cells to mate (Fig. 2c, d, e, Supplementary Fig. 3). The fact that the *W. anomalus* Matα2 does not regulate the **a**-specific genes can account for the observation that the homeodomain has acquired mutations that compromise its binding to the *S. cerevisiae* **a**-specific genes, as detected in our reporter assays. In addition, we analyzed the *W. anomalus* **a**-specific gene *cis*-regulatory elements bioinformatically and experimentally and found that they possess Mat**a**2-Mcm1, not Matα2-Mcm1 binding sites, entirely supporting the genetic experiments (Supplementary Fig. 4b). These results argue against the possibility that direct **a**-specific gene repression by Matα2 existed in an ancestor of *W. anomalus* but was subsequently lost, as this would have required the independent loss of Matα2 binding sites from all of the **a**-specific genes across numerous species.

The experiments described so far demonstrate that the new circuit—Matα2 repression of the **a**-specific genes—had partially formed by the time the *W. anomalus* clade branched but was not completed until after the divergence of the *S. cerevisiae* clade. The last step in the completion

of the circuit was the acquisition of *cis*-regulatory sequences recognized by Matα2 in the **a**-specific genes. This step must have occurred long after Matα2 gained the regions that bind Tup1 and Mcm1. We next address how these changes in the Matα2 protein could have predated the completion of the circuit.

One hypothesis for how Matα2 acquired the two protein-protein interactions needed to repress the **a**-specific genes millions of years before these genes acquired the necessary *cis*-regulatory sequences holds that these interactions are needed for Matα2's more ancient function in repressing the haploid-specific genes, but only in the *W. anomalus* clade. To test this idea, we analyzed the requirements for haploid-specific gene repression in *W. anomalus*. We deleted *MATα2* and *MAT**a**1* in **a**/α cells and found that they are both necessary for haploid-specific gene repression (Fig. 3a, b, Supplementary Fig. 5). This is the case for virtually all species in the three clades studied to date. However, unlike species outside the clade, we found using genetic experiments that regions 1 and 3 of *W. anomalus* Matα2 (the Tup1-interaction region and the Mcm1-interaction region) are necessary for repression of the haploid-specific genes (Fig. 3c). Finally, we found that a Mcm1 *cis*-regulatory site is also required for the repression of the haploid-specific gene *RME1* (Fig. 3d, e, Supplementary Fig. 6). Taken together with the results of chromatin immunoprecipitation experiments, we show that Matα2, Mat**a**1, and Mcm1 are all required for haploid-specific gene repression in *W. anomalus* and that the portions of Matα2 that interact with Mcm1 and Tup1 are also required. This three-part recognition of the haploid-specific genes in the *W. anomalus* clade was not anticipated from studies of other species. Even in the *S. cerevisiae* clade where Mcm1 and Matα2 are known to interact, this interaction is not required for haploid-specific gene repression. This analysis explains the initially puzzling observation that the key changes in Matα2 that were needed to form the new **a**-specific gene circuit were already in

place in the last common ancestor of *S. cerevisiae* and *W. anomalus,* long before the circuit was completed (Fig. 4).

Discussion

In this paper we examine the evolutionary trajectory through which a new transcriptional circuit was formed in a group of fungal species that includes *S. cerevisiae*. The key regulatory protein responsible for this new circuit (the homeodomain protein Matα2) is ancient, and the genes it controls in the new circuit (the **a**-specific genes) are also ancient. We show that the new, derived circuit—which links Matα2 to the **a**-specific genes by direct binding interactions—formed in at least two stages, separated by millions of years. In the first stage, Matα2 acquired two short amino acid sequences, each of which recognizes an additional ancient protein (Tup1 and Mcm1). In the second stage, the **a**-specific genes acquired *cis*-regulatory sequences recognized by Matα2. The changes in the coding sequence of Matα2 are observed across a broad clade of species, but the *cis*-regulatory changes are seen only in a subset of those species.

This scenario raises the question of how the modifications of Matα2 could arise and be maintained long before the circuit was completed by the *cis*-regulatory changes in the **a**-specific genes. By analyzing a group of species that branched after the modifications in Matα2 had occurred but before the *cis*-regulatory changes were in place, we found the modifications of Matα2 were now needed for Matα2 to carry out its deeply conserved ancestral function. This clade-specific requirement explains how the modifications of Matα2 could be maintained for long periods of evolutionary time before the new **a**-specific gene repression circuit was completed.

This study helps to illuminate several long-standing issues. (1) How is pleiotropy avoided when transcriptional regulators acquire new functions? The modular structure of Matα2 is evident from experiments showing that regions of the protein (Tup1 and Mcm1-interaction regions) that were acquired recently can be transplanted to a variety of outgroup Matα2 proteins (representing

the ancestral Matα2) and that they endow the ancestral proteins with the new functions without compromising the existing functions (Baker et al., 2012). Matα2 therefore can "cleanly" acquire new functions without compromising existing functions. Virtually all eukaryotic transcriptional regulators have analogous modular structures, and pleiotropy (e.g. off-target effects) can be avoided in the same way. But there is a second, more subtle way that pleiotropy was avoided in the case studied here. The constraints of Matα2's deeply conserved, ancestral function meant that only certain evolutionary trajectories were available to the protein. Its ancestral function requires of Matα2 to work in combination with Mat**a**1 to bind and repress the haploid-specific genes. Our work shows that there are at least two ways that this binding can occur. It can happen solely through a Mat**a**1-Matα2 heterodimer (as it does the *S. cerevisiae* clade) or it can occur through a triple combination with Mat**a**1, Matα2 and Mcm1 (as is found in *W. anomalus*). All three of these proteins are ancient, and as a default hypothesis, we propose that this shift occurred neutrally with the energy of binding to the control region of the haploid-specific genes simply being "shared out" differently in the two groups of species. According to this idea, the newly acquired interaction between Matα2 and Mcm1 could have occurred only if it was compatible with Matα2's existing function. In this way, pleiotropy was avoided almost automatically; even before the new **a**-specific gene circuit was formed, the Matα2-Mcm1 combination (which forms the basis of the new circuit) had been "vetted" for millions of years as being compatible with the ancestral functions of Matα2.

(2) Is this evolutionary scenario compatible with the concept of "constructive neutral evolution," the idea that new functions can evolve through evolutionary transitions of approximately equal fitness (Gray et al., 2010; Lukeš et al., 2011; Stoltzfus, 1999)? Constructive neutral evolution means that new functions need not be dependent on ever-improving performance; rather, complexity can be initially gained neutrally, providing an array of different "solutions" to a regulatory problem. Before the results presented here, it was difficult to understand

how the derived circuit represented by *S. cerevisiae* (repression of the **a**-specific genes by Matα2 in α cells) could have evolved because it required changes in both the Matα2 coding region and in the *cis*-regulatory sequences controlling the 5-10 **a**-specific genes. We previously speculated (incorrectly, as it turned out) that the *cis*-regulatory changes in the **a**-specific genes occurred before the coding changes in the Matα2 protein (Baker et al., 2012). The present study rules out that hypothesis by showing that the coding-sequence changes must have come first and that they are needed in the extant *W. anomalus* clade for Matα2 to carry out its deeply conserved, ancestral function. We propose that this change represents an example of constructive neutral evolution, in the sense that the neutral sampling of different ways to repress the haploid-specific genes over evolutionary time led to changes in Matα2 that, millions of years later, through exaptation, formed the basis of the new circuit. Without this exploration of different molecular solutions to repress the haploid-specific genes it is difficult to imagine how the new **a**-specific gene circuit could have formed, as it would require numerous coding changes and *cis*-regulatory changes happening more or less simultaneously. Although we cannot rule out the possibility that the changes in the way that the haploid-specific genes were repressed were somehow adaptive, it seems more likely that they occurred neutrally, an explanation consistent with a wide variety of theoretical work (Gray et al., 2010; Lukeš et al., 2011; Stoltzfus, 1999; Wagner, 2005). In any case, there is no obvious adaptive explanation and neutral evolution is an appropriate default hypothesis.

(3) Is there a "logic" to the structure of transcription circuits? Over the past ten years we have documented many changes in the regulation of the cell-type specific genes (the **a**-specific genes, the α-specific genes, and the haploid-specific genes) across a broad fungal lineage spanning nominally 300 million years. We have observed changes in the DNA-binding specificity of regulatory proteins, the making and breaking of combinations of regulatory proteins, gains and losses of *cis*-regulatory sequences, changes from positive to negative control, changes from direct

to indirect regulation, and examples of circuit epistasis. In this paper we show that some clades regulate the haploid-specific genes with a combination of three proteins, while other use only two of the proteins. Throughout all of these changes which have taken place in different lineages, the overall pattern of cell type specific genes expression has largely remained the same: **a**-specific genes are expressed in **a** cells, α-specific genes in α cells, haploid-specific genes in **a** and α cells and all three classes of genes are off in **a**/α cells. If there is any overriding "design logic" to the different mechanisms of regulating these genes, it is difficult to discern. Rather, the spectrum of observed changes seem to reflect a general principle, clearly illustrated by this case: changes in regulatory proteins and *cis*-regulatory sequences that do not compromise existing circuits will be favored. The evolutionary trajectories that are subject to this constraint can provide a source of novelty that can account for the widespread variety in the regulatory mechanisms underlying cell-type specification in fungi. More broadly, the work presented here illustrates that a given transcription circuit is best understood as one of several possible interchangeable solutions rather than as a finished, optimized design (Sorrells and Johnson, 2015).

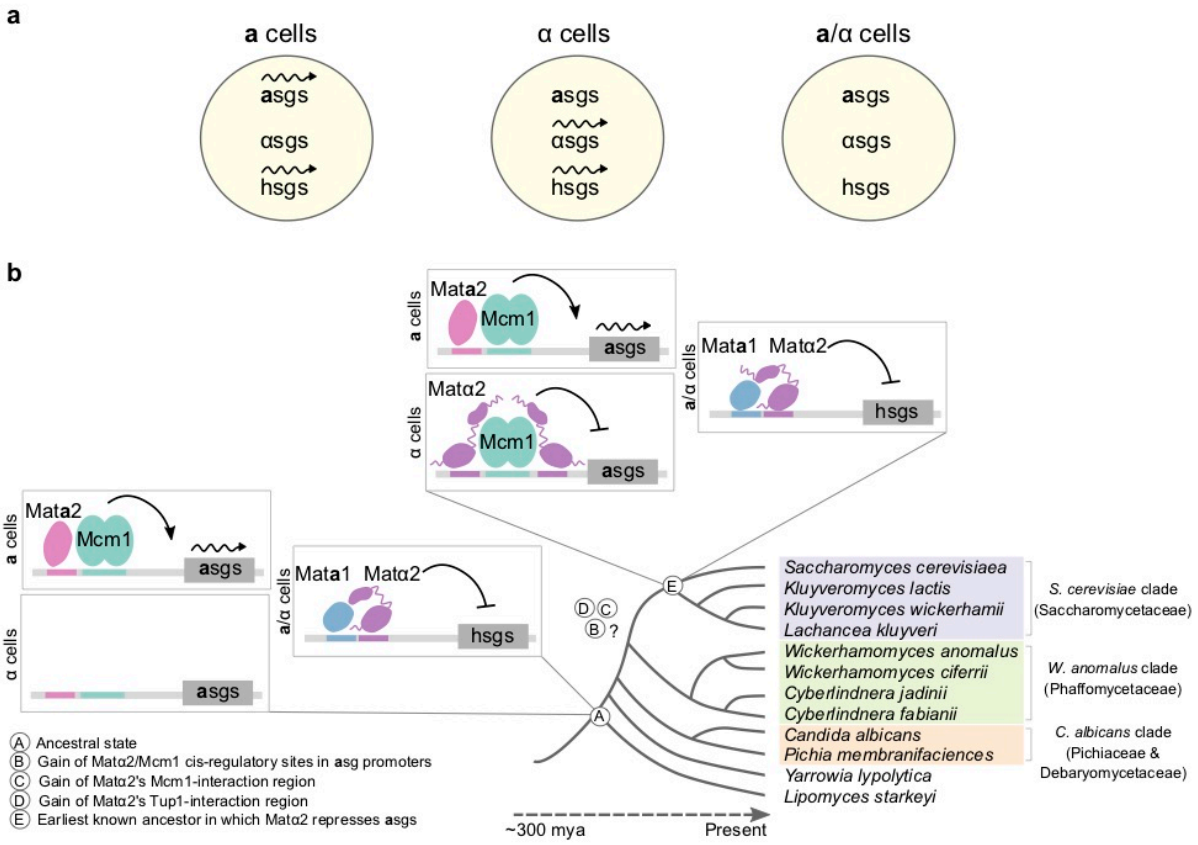
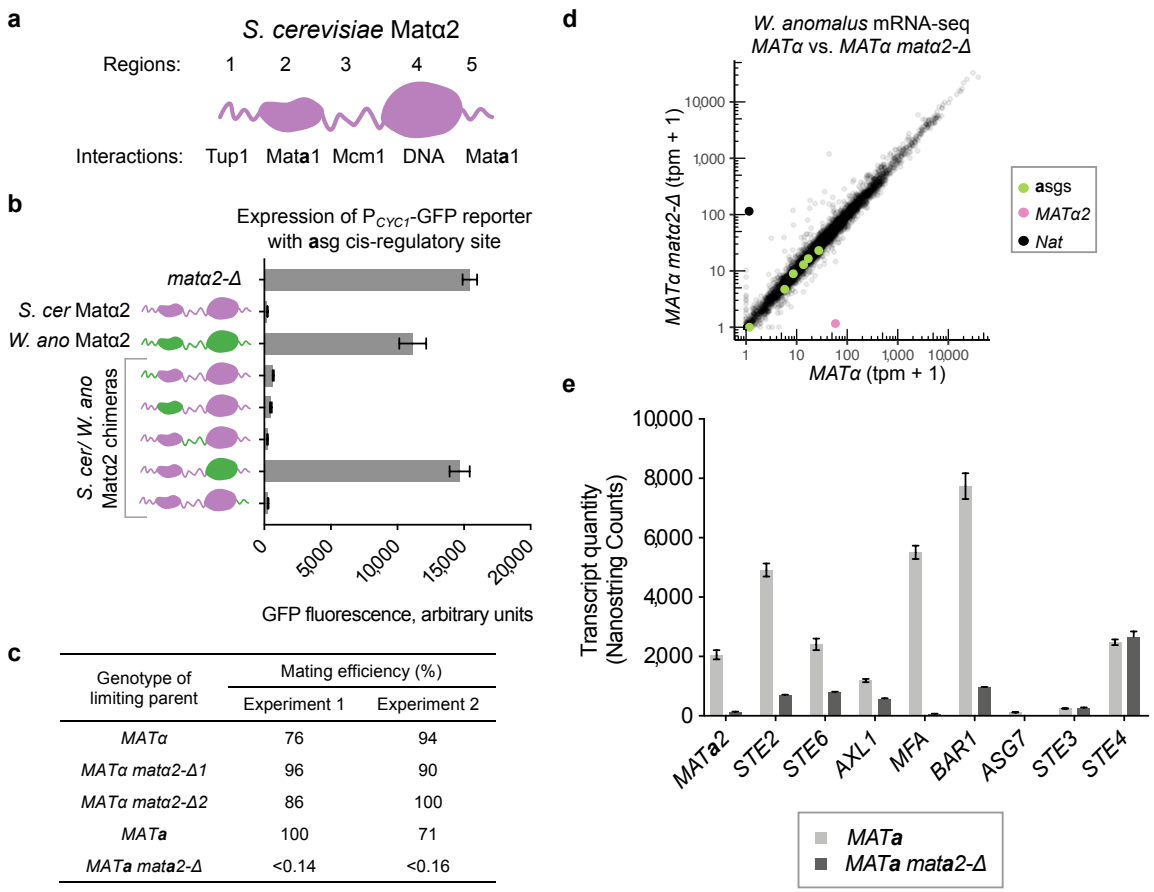


Figure 1. Cell-type specific gene expression in the Saccharomycotina yeast

a. Across the entire Saccharomycotina clade, **a** and α cells each express a set of genes specific to that cell type (**a**- and α-specific genes, or **a**sgs and αsgs, respectively), as well as a shared set of haploid-specific genes (hsgs). **a** and α cells can mate to form **a**/α cells that do not express the **a**-, α- or haploid-specific genes (Herskowitz, 1989).

b. The mechanism underlying the expression pattern of **a**-specific genes is different among species. In the last common ancestor of the Saccharomycotina yeast (circled A in the figure) transcription of the **a**-specific genes was activated by Mat**a**2, an HMG-domain protein only produced in **a** cells, which binds directly to the regulatory of each **a**-specific gene (Tsong et al., 2003a; 2006). Much later in evolutionary time (E), repression of the **a**-specific genes by direct binding by Matα2 evolved. Steps B, C, and D must have occurred between steps A and E. Their order is addressed in this paper. As indicated in the schematic diagrams, both Matα2 and Mat**a**2 bind DNA with Mcm1, a transcriptional regulator found in all cell types. The schematic diagrams are based on a large body of published literature including genetic, biochemical, and structural studies.



Genotype of limiting parent	Mating efficiency (%)	
	Experiment 1	Experiment 2
MATα	76	94
MATα matα2-Δ1	96	90
MATα matα2-Δ2	86	100
MATa	100	71
MATa mata2-Δ	<0.14	<0.16

Figure 2. *W. anomalus* Matα2 has functional Tup1 and Mcm1-interacting regions but does not repress the a-specific genes

a. The five modules of the *S. cerevisiae* Matα2 protein. Structural domains are shown as globular, and unstructured regions are shown as wavy lines.

b. Activity of Matα2 derivatives in repressing **a**-specific gene transcription in *S. cerevisiae*. The **a**-specific gene *cis*-regulatory element from the *S. cerevisiae* **a**-specific gene *STE2* gene was placed into a reporter construct and used to measure the ability of *S. cerevisiae* Matα2 (purple), *W. anomalus* Matα2 (green), and hybrid proteins (purple and green) to repress **a**-specific gene transcription. *MATα2* constructs (driven by the *S. cerevisiae MATα2* promoter) were inserted into a *S. cerevisiae* α cell with *MATα2* deleted (*matα2-Δ*). In addition, regions 1-5 of the *W. anomalus* Matα2 (green) sequence were swapped into the *S. cerevisiae* Matα2 (purple) protein and tested for their ability to it. Mean and SD of three independent genetic isolates grown and tested in parallel are shown.

c. In *W. anomalus*, Mat**a**2 is required for **a** cells to mate but Matα2 is not required for α cells to mate. The results of quantitative mating assays (see Supplementary Information) are given for wild-type **a** and α cells and various mutant derivatives. The values show the percentage of cells of the limiting parent strain that mated in the experiment.

d. mRNA-seq of wildtype *W. anomalus* α cells (*MATα*) compared to α cells with *MATα2* deleted (*MATα2 matα2-Δ*). *MATα2* expression in transcripts per million (tpm) is shown in

pink, confirming it has been deleted. The **a**-specific genes *STE2*, *AXL1*, *ASG7*, *BAR1*, *STE6,* and *MAT**a**2* are shown in green. The marker used to delete *MATα2* (*Nat*) is shown in opaque black (providing an additional control for the experiment), and all other mRNAs are shown in translucent black. Data from one culture of each genotype is plotted here, and data from independent replicates grown and prepared in parallel and from an independent deletion of *MATα2* are given in Supplementary Fig. 3.

e. **a**-specific gene expression levels in a wildtype *W. anomalus* **a** cells (*MAT**a***) vs. **a** cells with *MAT**a**2* deleted (*MAT**a**2 mat**a**2-Δ*), measured by the NanoString nCounter system (Geiss et al., 2008). All **a**-specific genes require *MAT**a**2* for wildtype-level expression in the **a** cell. Also shown are expression levels of the α-specific gene *STE3* and the haploid-specific gene *STE4*. Mean and SD of two cultures per genotype, grown and tested in parallel, are shown.

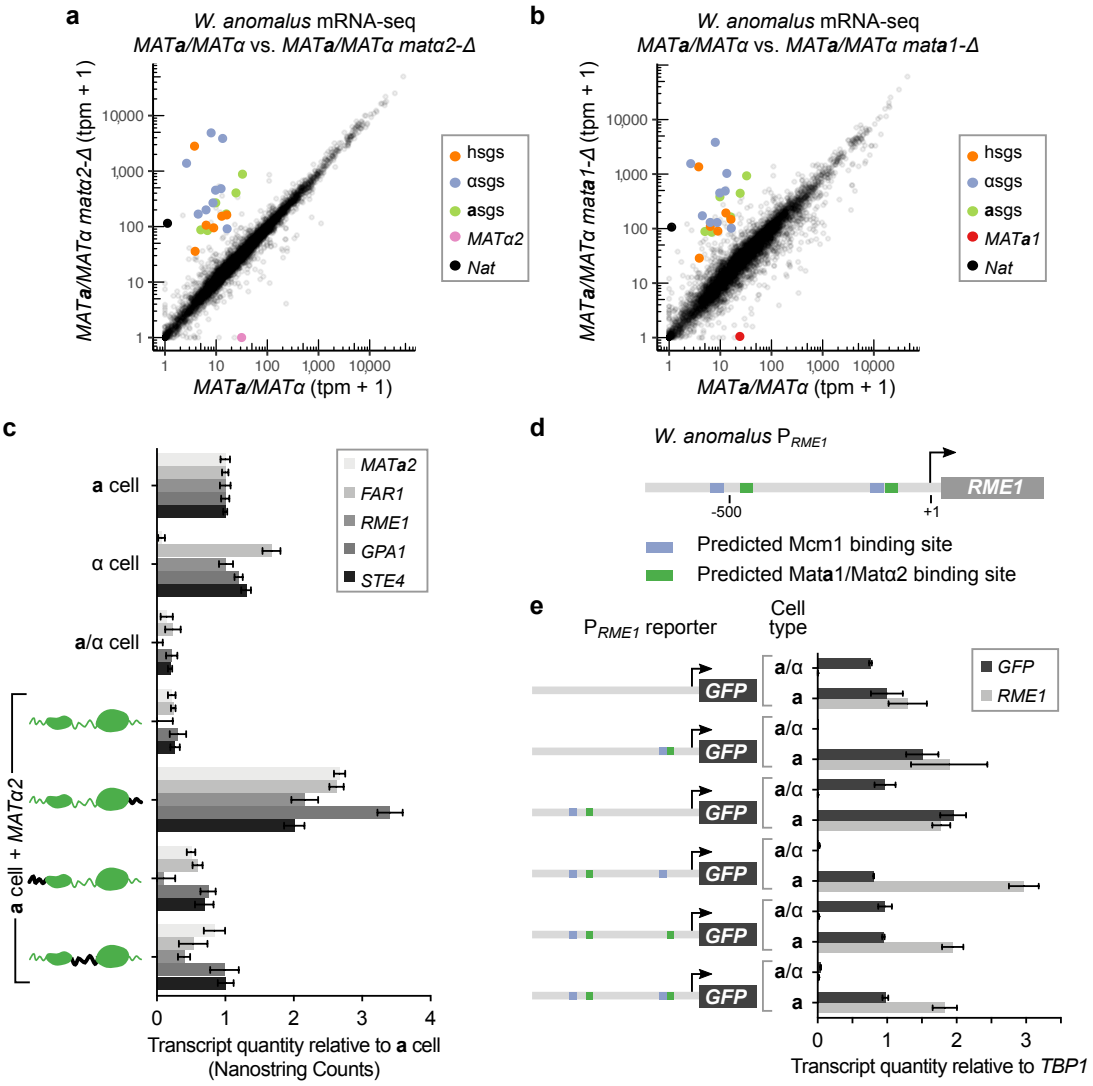


Figure 3. Mata1, Matα2, and a Mcm1 *cis*-regulatory sequence are all required for haploid-specific gene repression in *W. anomalus*

a. Matα2 is required for haploid-specific gene repression. mRNA-seq of wildtype *W. anomalus* **a**/α cell (*MAT**a**/MATα*) compared to an **a**/α cell with *MATα2* deleted (*MAT**a**/MATα matα2-Δ)*. As in the previous figure, *MATα2* expression in transcripts per million (tpm) is shown in pink, confirming it has been deleted. The **a**-specific genes *STE2*, *AXL1*, *ASG7*, *BAR1*, *STE6,* and *MAT**a**2* are shown in green, the haploid-specific genes *STE4*, *GPA1*, *FUS3*, *SST2*, *RME1*, and *FAR1* are shown in orange, the α-specific genes *SAG1*, *STE3*, *STE13*, *MFα* (multiple copies), *AFB1* (multiple copies), and *MATα1* are shown in blue-gray. The marker used to delete *MATα2* (*Nat*) is shown in opaque black, and all other mRNAs are shown in translucent black. The results show that the **a**-, α- and haploid-specific genes all increase in expression when *MATα2* is deleted from an **a**/α cell. The **a**- and α-specific genes are expressed because the proteins that activate them (Mat**a**2 and Matα1) are normally repressed in an **a**/α cell but are expressed when *MATα2* is deleted. Data from one culture of each genotype is plotted here, and data from replicates grown and prepared in parallel are in Supplementary Fig. 5.

b. Mat**a**1 is required for haploid-specific gene repression in *W. anomalus*. mRNA-seq of wildtype *W. anomalus* **a**/α cells (*MAT**a**/MATα*) compared to **a**/α cells with *MAT**a**1* deleted (*MAT**a**/MATα mat**a**1-Δ)*. All genes are labeled as in previous panel, except *MAT**a**1* is shown in red, confirming that it has been deleted.

c. Diagram of the sequence upstream of the *RME1* coding sequence indicating presumptive Mat**a**1/Matα2 (green) and Mcm1 (blue) binding sites.

d. Expression levels of endogenous *RME1* transcript (which serves as a control) and various P_{RME1}-GFP reporter constructs in *W. anomalus* **a** and **a**/α cells measured by RT-qPCR. As indicated, a variety of different deletions were made in the reporter. The results show that the predicted Mcm1-binding site nearest to the transcription start site is necessary for *RME1* repression in the **a**/α cell. Quantities are mean and SD of two cultures grown and measured in parallel, normalized to expression of the housekeeping gene *TBP1*. The experiment was repeated with independent genetics isolates of each strain and each repeat recapitulated the results (Supplementary Fig. 6).

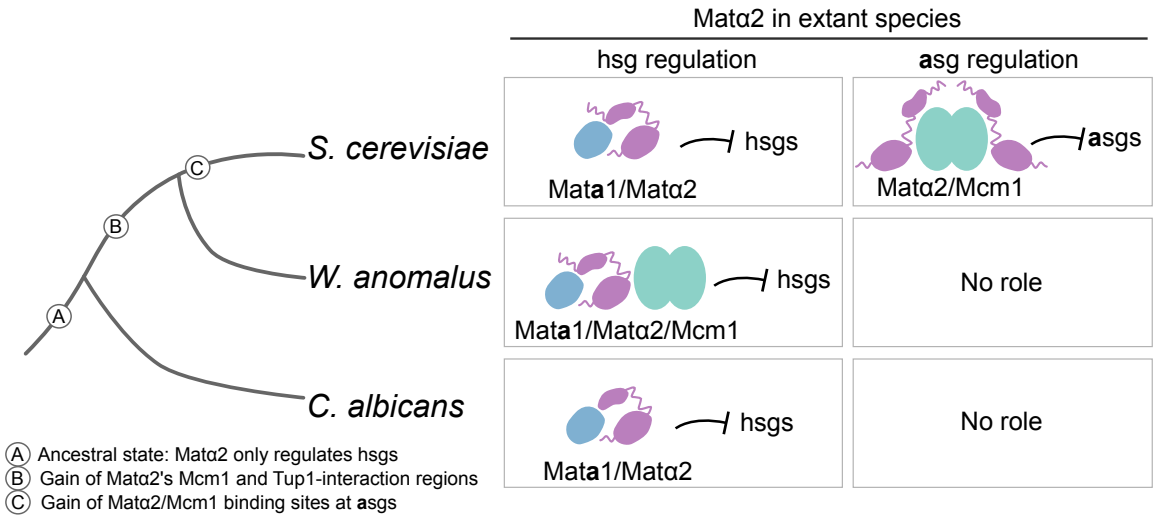


Figure 4. Order of evolutionary events leading to repression of the a-specific genes by Matα2, mapped on to the yeast cladogram

Shown at the right is the role of Matα2 in three extant species. The three-protein solution for repressing the haploid-specific genes is still in place in the *W. anomalus* clade, but in the *S. cerevisiae* lineage was partitioned into **a**-specific gene regulation (which uses only two proteins, Mcm1 and Matα2) and repression of the haploid-specific genes (which requires Matα2 and Mat**a**1). Without the "three-protein" intermediate it is difficult to imagine how **a**-specific gene regulation could have evolved. In particular, the intermediate explains how the necessary changes in the regulatory protein Matα2 could be maintained for millions of years before the new **a**-specific regulatory circuit was completed by acquisition of *cis*-regulatory sequences.

Supplementary Information

Yeast cell-type specific gene expression regulation

Saccharomycotina yeast have two mating types termed **a** and α cells that are genetically identical except for a short DNA segment called the mating-type locus (*MAT*) (Herskowitz, 1989). This segment codes for transcriptional regulators that ensure the proper expression of the **a**, α, and haploid-specific genes. Thus, **a** cells express the **a**-specific and haploid-specific genes (but not the α -specific genes), and α cells express the α-specific and haploid specific (but not the **a**-specific genes) (Fig. 1a). When an **a** cell successfully mates with an α cell, an **a**/α cell is formed. This cell type does not express the **a**, α, or haploid-specific genes but is competent to undergo meiosis. The proper regulation of these three sets of genes, which is directed by the regulatory proteins coded at the mating type locus, is thus essential not only for proper mating, but also for meiosis, which completes the fungal sexual life cycle.

W. anomalus Matα2 has functional Mcm1 and Tup1-interaction regions

Given the similarity of the *W. anomalus* Matα2 protein sequence to that of *S. cerevisiae*, we hypothesized that the *W. anomalus* protein represented the derived form of Matα2. To test this hypothesis, we expressed the *W. anomalus* and, as a control, the *S. cerevisiae MATα2* coding sequences in a *S. cerevisiae* α cell that had the *MATα2* gene deleted and contained a P_{CYC1}-GFP reporter with an **a**-specific gene *cis*-regulatory sequence taken from the promoter of the *S. cerevisiae* **a**-specific gene *STE2*. As expected, the *S. cerevisiae* Matα2 represses transcription from this promoter by more than 99 percent. In contrast to our expectations, *W. anomalus* Matα2 failed to repress the reporter (Fig. 2b). By constructing a series of chimeric proteins, we mapped the "deficiency" of the *W. anomalus* Matα2 to its homeodomain, indicating that although the protein is in large part conserved, its homeodomain simply did not recognize the *S. cerevisiae* **a**-specific gene *cis*-regulatory sequence (Fig. 2b). To test this idea, we swapped the *S. cerevisiae*

homeodomain into *W. anomalus* Matα2 and we found that it could now carry out repression of the reporter construct, although not as well as the bona fide *S. cerevisiae* protein (Supplementary Fig. 1d). *W. ciferrii* and *C. jadinii* Matα2 regions 1 and 2 also resemble the *S. cerevisiae* sequence, and their full-length sequences and the *C. jadinii* homeodomain also do not support **a**-specific gene repression (Supplementary Fig. 1d)

W. anomalus a**-specific genes are not directly regulated by Matα2**

To determine whether the *W. anomalus* **a**-specific genes are repressed by Matα2, activated Mat**a**2, or both repressed by Matα2 and activated by Mat**a**2, we deleted *MATα2* from an α cell and assayed gene expression changes genome wide by mRNA-seq. None of the **a**-specific genes increased in expression (Fig. 2d, Supplementary Fig. 3), indicating that the *W. anomalus* Matα2 does not directly repress the **a**-specific genes. We further tested this conclusion using a more classical approach. In *S. cerevisiae*, *MATα2* is necessary for proper mating of α cells; its deletion causes ectopic expression of the **a**-specific genes thereby preventing efficient mating. In *W. anomalus* a wildtype α cell and two independently derived deletions of *MATα2* from α cells all mated with similar efficiency, indicating, unlike in *S. cerevisiae*, that *MATα2* is not required for α cell mating (Fig. 2c). These experiments provide additional support for the conclusion that the *W. anomalus* Matα2 does not regulate the **a**-specific genes in *W. anomalus*.

W. anomalus a**-specific genes are activated by Mat**a**2, the ancestral mode of gene expression**

The fact that *W. anomalus* Matα2 does not directly repress the **a**-sepecific genes suggests that, by default, the transcriptional activator Mat**a**2 must positively regulate the **a**-specific genes in *W. anomalus*. To test this hypothesis, we deleted *MAT**a**2* in an **a** cell and measured the expression levels of the **a**-specific genes with the NanoString nCounter system, a method for directly counting individual transcripts of choice in an RNA sample (Geiss et al., 2008). Each of the **a**-specific genes was significantly under-expressed in the *MAT**a**2* deletion strain with respect

to a wildtype **a** cell (Fig. 2e). We also tested the consequences of *MAT**a**2* deletion on **a**-specific mating in a quantitative mating assay and found that an **a** cell with *MAT**a**2* deleted fails to mate at any detectable level in our assay (Fig. 2c). Thus, *W. anomalus* **a**-specific genes are directly regulated by Mat**a**2-activation alone, and Matα2 repression is not involved. Consistent with this conclusion are bioinformatic and experimental analyses of the **a**-specific genes in *W. anomalus, C. jadinii and W. ciferrii,* which identifies *cis*-regulatory sequences that are similar to those controlled by Mat**a**2 in other species (Supplementary Fig. 4a, b). As discussed above, this scheme of regulating the **a**-specific genes is also observed in *C. albicans* and other outgroup species and, by inference, is likely to represent the situation in the last common ancestor of the three clades.

***W. anomalus* Matα2 retains the deep ancestral function of repressing the haploid-specific genes with Mat**a**1**

Matα2's more ancient function is repression of the haploid-specific genes with Mat**a**1. To test whether this ancestral role is retained in *W. anomalus*, we separately deleted either *MATα2* or *MAT**a**1* in an **a**/α cell background and measured gene expression changes by mRNA-seq. In both deletion strains the haploid-specific genes are significantly de-repressed compared to a wildtype **a**/α cell (Fig. 3a, b, Supplementary Fig. 5). We note that deletion of either *MAT**a**1* or *MATα2* also de-represses the *MAT**a**2* and *MATα1* genes, and the **a** and α specific genes are also improperly expressed in the deletion strains. This result also explains how the **a** and α-specific genes are kept off in the **a**/α cells. From these results, we conclude that *W. anomalus* retains the deep ancestral function of Matα2, to repress the haploid-specific genes in combination with Mat**a**1.

Mcm1 is necessary for haploid-specific gene repression in *W. anomalus*

We introduced a series of *MATα2* constructs into an **a** cell and measured changes in haploid-specific gene expression using the NanoString nCounter system. As expected, addition of a wildtype allele of *MATα2* resulted in transcriptional repression of the haploid-specific genes

relative to their expression levels in an **a** cell (Fig. 2c). *RME1*, a deeply conserved haploid-specific gene, was especially tightly repressed, to a level less than one percent of that in **a** cells. Addition of a *MATα2* allele with a point mutation in the Mat**a**1-binding region (region 5), on the other hand, failed to repress *RME1* and the other haploid-specific genes. Addition of *MATα2* chimeras where the Tup1 or Mcm1 interaction regions of Matα2 were replaced by the homologous region of the *C. albicans* (outgroup) Matα2 resulted in substantial loss of repression of many haploid-specific genes. We conclude that, in contrast to the case in *S. cerevisiae*, both the Tup1 and Mcm1-interacting regions of Matα2 are necessary for the complete repression of haploid-specific genes in *W. anomalus.*

The requirement for Matα2 to carry the Mcm1-interaction region in order to repress the haploid-specific genes in *W. anomalus* was surprising because this region is dispensable for haploid-specific gene repression in *S. cerevisiae* and *C. albicans* (Baker et al., 2012). We therefore carried out an experiment to test whether Mcm1 has a direct role in repressing the haploid-specific genes in *W. anomalus*. Mcm1 controls the expression of hundreds of genes and is an essential gene in all fungal species tested to date (Tuch et al., 2008). Therefore, experiments that alter levels of Mcm1 are difficult to rigorously interpret. Instead, we turned to the regulatory region of the *RME1* gene, the most tightly regulated haploid-specific gene, and found two high-scoring matches to the Mcm1 recognition sequence (a *cis*-regulatory sequence that is virtually invariant across these species) (Tuch et al., 2008) (Fig. 3d, Supplementary Fig. 6). We confirmed that both Mat**a**1 and Matα2 bind at the *RME1* locus by chromatin-immunoprecipitation followed by qPCR (Supplementary Fig. 6b). To test whether Mcm1 is also required for *RME1* repression in an **a**/α cell we constructed a reporter driven by the DNA upstream of the *RME1* gene, showed that it was repressed in **a**/α cells, and tested the effect of deleting the Mcm1 recognition sequences. This

experiment has the advantage of maintaining proper regulation of all the normal haploid specific genes, including the endogenous copies of *RME1*.

Deletion of the proximal, but not the distal Mcm1 recognition sequence did away with repression of the *RME1* reporter in **a**/α cells (Fig. 3e, Supplementary Fig. 6c). Based on motif searching, we also identified several potential Mat**a**1/Matα2 binding sequences in the *RME1* control region (Supplementary Fig. 6a). While the two best matches to the Mat**a**1/Matα2 binding site were not required for *RME1* repression, deletion of a possible Matα2 site that abuts the critical Mcm1 site also de-represses *RME1* in an **a**/α cell (Supplementary Fig. 6d). From these results we conclude that Mat**a**1, Matα2, and Mcm1 are all required to repress *RME1* transcription in **a**/α cells. Also required are the regions of Matα2 that bind Mcm1 and Mat**a**1.

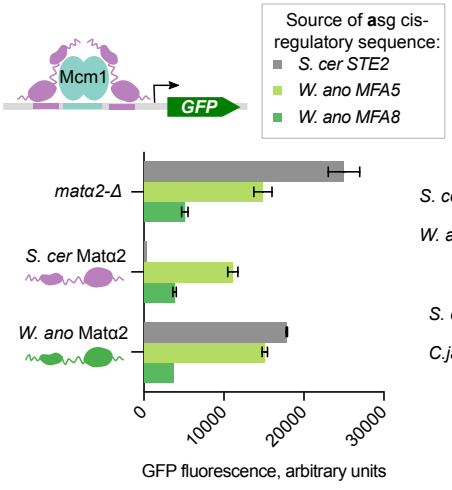
a

Species	Matα2 Region 1 (Tup1-interaction)	Matα2 Region 3 (Mcm1-interaction)
Lsta	----------MFYNTDLAISNDSN	RIAQRLALPDPIVGTSF
Tcas	MSSPSSNIPKESEGSRSTSCDSQPCIGTSEILNESHFEYFSDE	KIRTCDQDEDQIDEYQR
Ylip	----------MKETRDLLFYNENSIDFNFSPEEMDTHQMM	GQLLDLCKEGLQLVESI
Ptan	----------MAPQRNDRIYDKCSIPSLLVETPENQNQKL	EMVFNVITQNMINEFGS
Ppas	----------	--DVEIILQQILDSNST
Opol	----------MAFSCVAIPLLTPSQFEGCSNCHHDRS	SSMEGFTTERSLKYTRD
Bbru	----------MTRITTSK----------N	QNTYGSCKSNFIVTNGY
Pmem	----------	RIDFEDKIIPHTLSTQS
Hbur	----------MKVDNYQ-------PKQD	KISFEVLTQSDISK---
Stan	----------	AIGLNESGQTTQSNSHL
Spas	----------	SLGMLREEAVQEGADS
Cdub	----------	DSSFSNTVDDAIEDGDD
Calb	----------	DSPFSNSADTIDEDDDK
Ctro	----------	TKEEELKKEENLESDPS
Arub	----------MNNYSVDESDLASKIIVLSQN	NFKFAVVITPSEVRKKR
Wcif	----------MENSPISVHEKCSISSLLSEP---SSGLS	NLIFNIVTPSMISENGA
Wano	----------MDQSPNAIQEKCSISSLLTESNITNRTLS	RLVFNVITQRMMRESSN
Cjad	----------MENSSIQDKCSITSLLSDG---SKRRD	SMSFKIVTQDMMNKNGS
Ecym	----------MNRIPIISLLNPSPGDSHNIR	GIIFNVVTQDMMENGRK
Kmar	----------MSKIPVQSLLNPESPREHFYK	TVEFRILSSEHLNPTLN
Klac	----------MSRIPIHSLLNPSESCKSISN	SVNIKILTQDLMHSNNN
Zrou	----------MNKIPIERLLNPV-------H	GMVFNIVTQDTLNSARR
Tbla	----------MNKIPIDTLLNPT-------N	PRVFNVLTQHMMSNSNS
Vpol	----------MNKIPINILLNPS-------H	EFVFNIVTQDMMNKKKM
Tpha	----------MNKIPINKLLNPS-------Q	ILVFNVITQDMMNNNKK
Ncas	----------MNKIAIHDLLNPP-------N	KLVFNILTKKRMSPTSN
Ndai	----------MNKIAISDLLNPP-------T	KYVFNVITQDMMIPEKN
Kafr	----------MNKIQIQDLLNSP-------R	KTVFNVVTQDMMNLNRT
Knag	----------MNRIEIQDLLNNQ-------D	EFVFSVVTQDMMNKRKE
Cgla	----------MSKKSRISITHLLNPI-------Q	EIVFNVVTKDHLLKKGV
Skud	----------MNKIPIKDLLNPQ-------I	GLVFNVVTQDMINKSTK
Smik	----------MFSAILQISSMLVFL-------L	GLVFNVVTQDMINKSTK
Scer	----------**MNKIPIKDLL**NPQ-------I	G**LVFNVVTQ**DMINKSTK

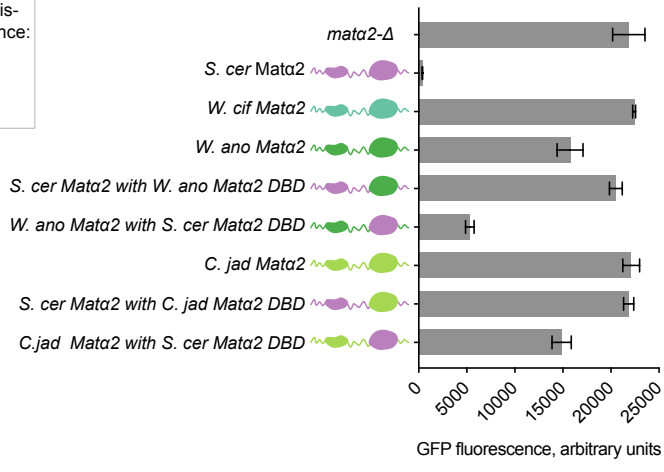
b

Species	Matα2 Region 4 (Homeodomain)
Lsta	NRSG--------ISKSS-KLQLRNWMHEHRRFPNLTAEEERDLQKTTGINKRQLDRW----------LTILRR--------RSL
Tcas	EDRP------NRLLKSKAVDVLSEWFEQNIDRPYPSMEVKQRLAKLAGCSVVQVNNW----------MGNARK--------RNF
Ylip	QHGGADAIDSEKLRKVKKLSELWEWFNAIPYNHHPSVSQIEFIASQIGEKVSFVKSW----------VANKRRTGAKKRSKKPA
Ptan	PNKG------RRLPKSL-VQILETWYEENIESPYLSEETLTLLVERTGLTKNQVRNW----------ISNKRR--------KEK
Ppas	-ERK------RRLPREK-VRILENWFADHFQHPYATQQDFDELVHLTSLSKTQIRNW----------ISNKRR--------KER
Opol	EKRS------KRFPKTA-QMELENWYTENEDNPYLSKRDLQQLVHKTGLCAPQVRNW----------VSNRRR--------KER
Bbru	RRKG------RRLSKPI-VSILSHWYEENIDNPYLTNSDIEELRAQTGLQDNQIRNW----------VSNRRR--------KK-
Pmem	PNHG------KRLPKET-VVILNEWYADNIDSPYLSKDDISYLMSKTHLRESQIKNW----------ASNKRR--------KKK
Hbur	GPKT------NRLQDHQ-VLILNEWYGRNKKHPYVDENSLKYLIEKTRLSETQVRNW----------FSNKRR--------KHR
Stan	-VSN------HKFSQNQ-VEILEQWYQSNSEHPFLNNQSTLELQTRTGLTVTQVKNW----------VSNKRR--------KEK
Spas	CKRG-----SRRLVKQQ-LDILENWYNQNLQHPYLTRESIIELMNLTSLSKSQVQNW----------ISNRRR--------KEK
Cdub	KTKS------RRLTKKQ-LLVLEGWFQKHKNHPYSQKDQTNLLIKSTKLSKSQVQNW----------ISNRRR--------KEK
Calb	KIKS------RRLTKKQ-LLVLEGWFQKHKNHPYSQKDQTNLLIKSTKLSKSQVQNW----------ISNRRR--------KEK
Ctro	KSKN------KRLSKNQ-LEILEGWFQKNKKHPYSQKDQTDLLMKSTNLSKSQVQNWYVLHFAFEVRDKNTNTLFRISNRRR--------KEK
Arub	PGSG------TSQKKYG-KKILEDWYHANSKYPYVGK-NIGRLCKLSGFDASKIRTW----------FSNKRR--------KDK
Wcif	NQKR------RRLESGK-VEALEDWYIKHTEKPYLNEETLNMLERMTGLTKVQIKNW----------ISNKRR--------KDK
Wano	PHRR------RRLERGK-VGVLQTWYNKHFEKPYLNEETLNSLVDLTGLTKVQIRNW----------ISNKRR--------KEK
Cjad	IGKR------KRLERSK-VAVLEQWYSKNCDRPYLSDSMLLSSMGETGLSKIQIKNW----------LSNKRR--------KER
Ecym	SYKG------HRLPKKH-VQLLESWYKNNIENPYLNDTDIRILMRETGFSRSQVKNW----------VANKRR--------KDK
Kmar	VGKK------NRFPKKS-IKLLENWYECNRNSPYLTSNHLDYLKCNTGLNKTQIKNW----------VANRRR--------KDK
Klac	FKKG------KRFPKSH-IQLLENWYSMNRRNPYLAENDLAYISKNTTLTKTQIKNW----------LANRRR--------KDK
Zrou	TYRG------HRLPKHI-TRLLESWFNRNIEHPYLQTSSMKELMVETKLSGPQIKNW----------VSNRRR--------KEK
Tbla	SHRG------HRLPKHT-LIPLEKWFLHNKSHPYLHNSDLQALTTQSSLSKTQVKNW----------ISNRRR--------KER
Vpol	SYRG------HRLPKQN-VKLLERWYIQNVENPYLDDKSISELMKITSLSKVQIKNW----------VSNRRR--------KEK
Tpha	SYRG------HRLPKKN-VQYLEDWYMDHRKNPYLNEINIKLLMSKTSLSRIQVKNW----------ISNRRR--------KEK
Ncas	PRRG------HRLAKEK-VDLLEHWYIQHMDNPYLNKASLQMLMQETSLSKMQIKNW----------VSNRRR--------KEK
Ndai	PHRG------HRLPKEK-VNRLEHWYLAHIQKPYLDSKNLKVLMEETKLSKVQIKNW----------ISNRRR--------KEK
Kafr	SYRG------HRFSKEN-VKVLERWYTAHIDRPYLNRQSTEYLISKTGLSRVQIKNW----------VSNRRR--------KEK
Knag	PCRG------HRFSSNS-TETLEDWYKKHHEKPYLDKRSLHELEFKTKLSKMQIRNW----------VSNRRR--------KEK
Cgla	SYRG------HRFTREN-VQILETWYRNHIDNPYLDHNSQQYLAQKTNLSKIQIKNW----------VANRRR--------KQK
Skud	PYRG------HRFTKEN-VRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNW----------VSNRRR--------KEK
Smik	PYRG------HRFTKEN-VRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNW----------VSNRRR--------KEK
Scer	PYRG------HRFTKEN-VRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNW----------VSNRRR--------KEK

c



d



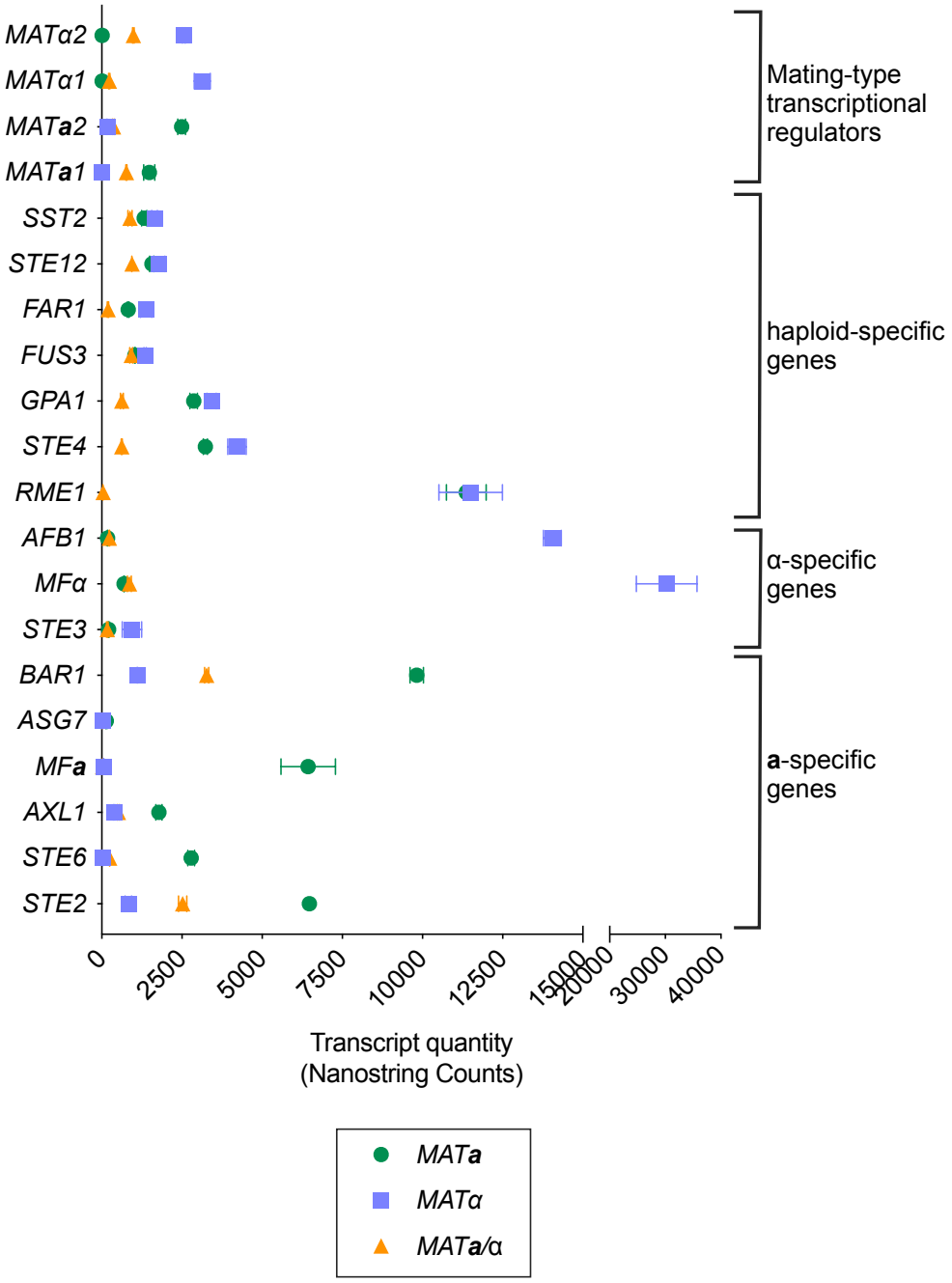
Supplementary Figure 1. *W. anomalus* clade Matα2 orthologs contain functional Tup1 and Mcm1-interaction regions, but cannot repress a-specific gene transcription due to changes in homeodomain

a. Multiple sequence alignment of Matα2 regions 1 and 3, generated by MUSCLE. *S. cerevisiae* clade (Saccharomycetaceae) sequences are shown in purple, *W. anomalus* clade (Phaffomycetaceae) sequences are shown in green, and *C. albicans* clade (Pichiacea and Debaryomycetaceae) sequences are shown in orange. Species are denoted with a four letter code where the first letter is the first letter of the genus name, and the next three letters are the first three letters of the species name (KOMACHI et al., 1994; Li et al., 1995; Mead et al., 1996; Tan and Richmond, 1998). Residues known to be functionally important for binding Tup1 and Mcm1 in *S. cerevisiae* are bolded on the *Scer* line.

b. Multiple sequence alignment of Matα2 region 4, the homeodomain, generated by MUSCLE. As above, *S. cerevisiae* clade (Saccharomycetaceae) sequences are shown in purple, *W. anomalus* clade (Phaffomycetaceae) sequences are shown in green, and *C. albicans* clade (Pichiacea and Debaryomycetaceae) sequences are shown in orange.

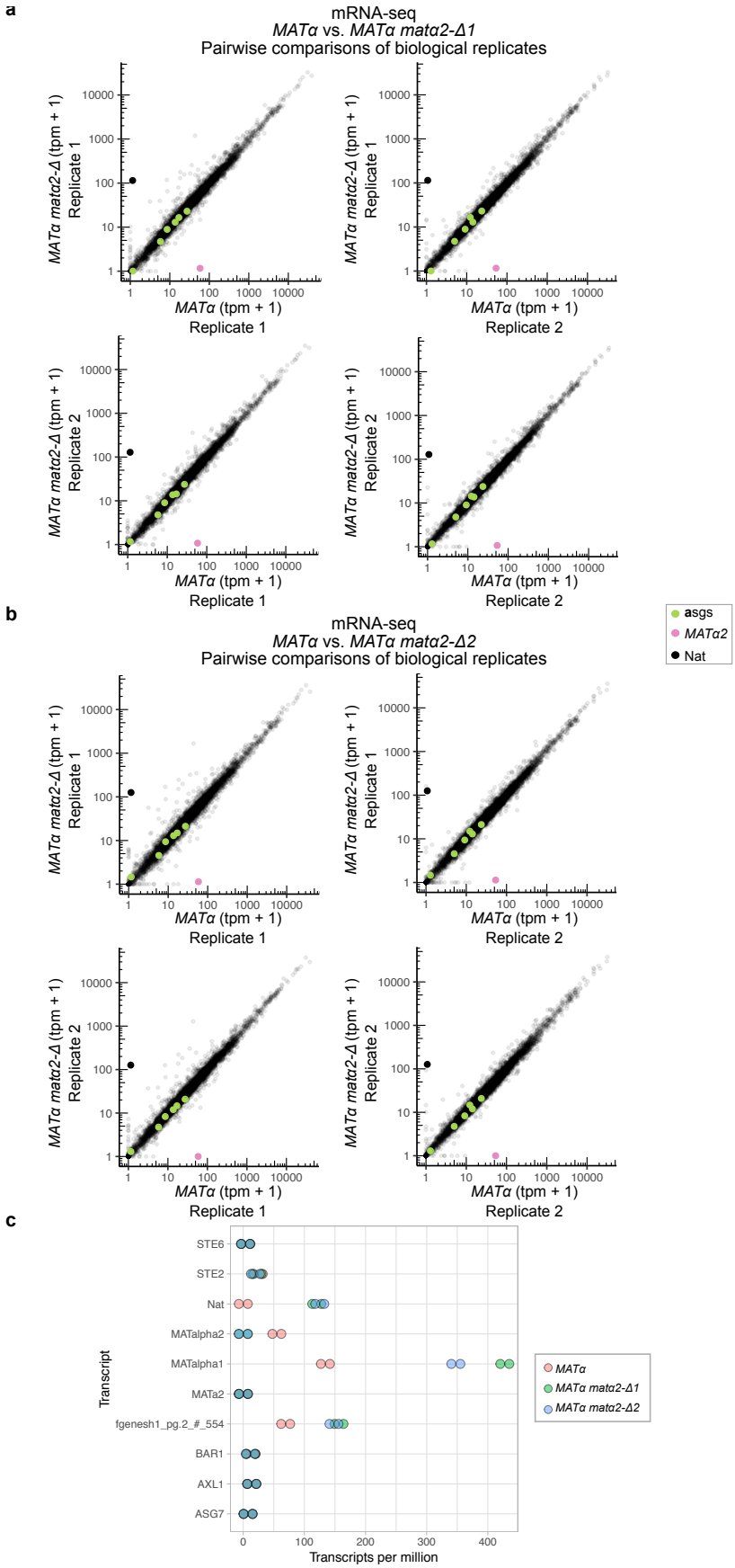
c. The **a**-specific gene *cis*-regulatory elements from the *S. cerevisiae STE2* gene, and two *W. anomalus MFA* genes were placed into a reporter construct and used to test the ability of *S. cerevisiae* Matα2 (purple) and *W. anomalus* Matα2 (green) to repress **a**-specific gene transcription. *MATα2* constructs were driven by the *S. cerevisiae MATα2* promoter and inserted into a *S. cerevisiae* α cell with *MATα2* deleted (*matα2-Δ*). Mean and SD of three independent isolates grown and tested in parallel are shown for all GFP reporter experiments.

d. *W. anomalus* clade Matα2 proteins, as well as chimeric Matα2 alleles with the *S. cerevisiae* Matα2 homeodeomain DNA-binding domain ("DBD") (purple) swapped into *W. anomalus* clade Matα2 protein sequences (*W. anomalus*, *W. ciferrii*, and *C. jadinii*) (shades of green), and vice versa, were tested for their ability to repress transcription of a reporter with the Matα2/Mcm1 binding site from *S. cerevisiae STE2* promoter. *W. anomalus* clade Matα2s consistently do not support **a**-specific gene transcriptional repression in this assay. For both *W. anomalus* and *C. jadinii*, this is due to the DNA-binding domain. Mean and SD of three independent genetic isolates grown and tested in parallel are shown.



Supplementary Figure 2. Cell-type specific gene expression patterns in *W. anomalus*

Cell-type specific gene expression levels in wildtype *W. anomalus* **a** (green circles), α (blue squares), and **a**/α (orange triangles) cells measured with the NanoString nCounter system. Mean and SD of two samples are plotted, except for cases in which the error bars are smaller than the data point. This is not intended to be an exhaustive list, and there may be additional cell-type specific genes.

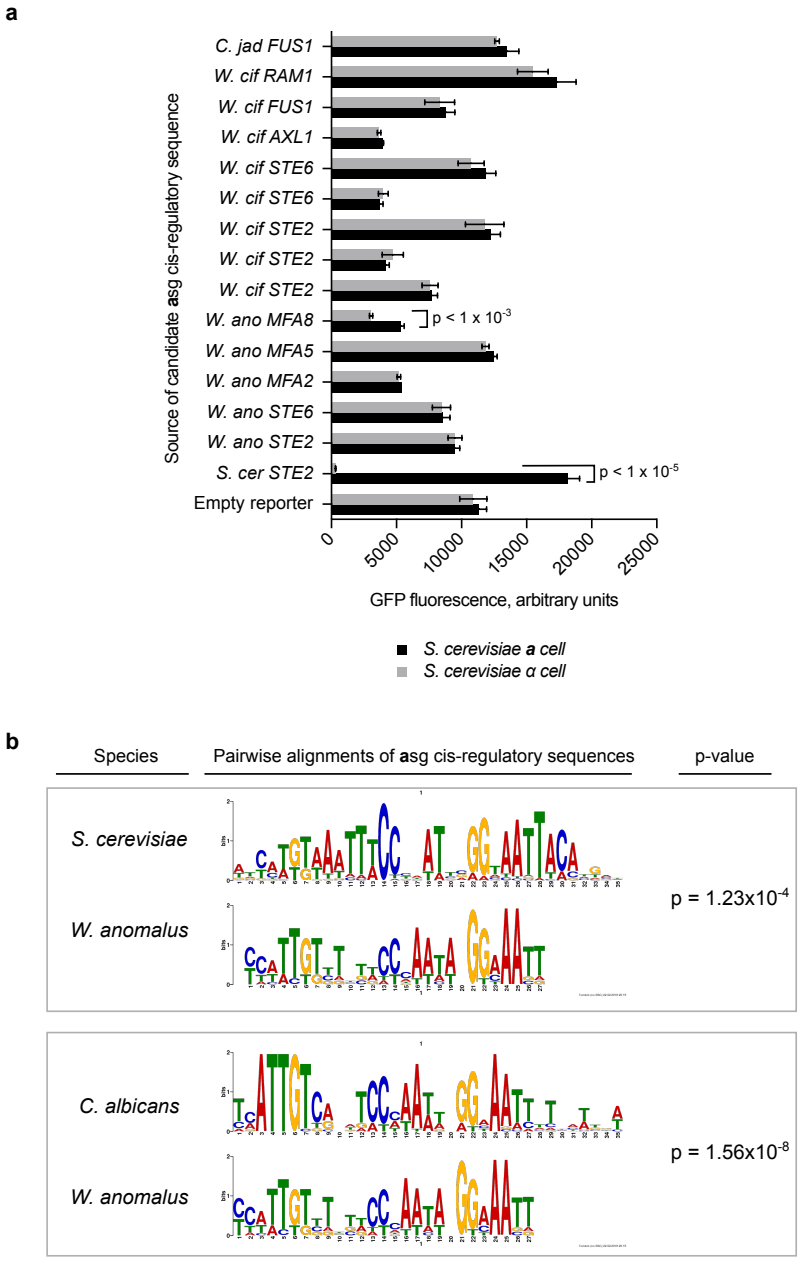

a
mRNA-seq
MATα vs. MATα matα2-Δ1
Pairwise comparisons of biological replicates
MATα matα2-Δ (tpm + 1) Replicate 1
MATα matα2-Δ (tpm + 1) Replicate 2
MATα (tpm + 1) Replicate 1
MATα (tpm + 1) Replicate 2
b
mRNA-seq
MATα vs. MATα matα2-Δ2
Pairwise comparisons of biological replicates
asgs
MATα2
Nat
c
STE6
STE2
Nat
MATalpha2
MATalpha1
MATa2
fgenesh1_pg.2_#_554
BAR1
AXL1
ASG7
Transcript
Transcripts per million
MATα
MATα matα2-Δ1
MATα matα2-Δ2

Supplementary Figure 3. *W. anomalus MATα2* does not directly repress the a-specific genes

a. mRNA-seq of wildtype *W. anomalus* α cells (*MATα*) vs. first isolate of α cells with *MATα2* deleted (*MATα2 matα2-Δ1)*. *MATα2* expression in transcripts per million (tpm) is shown in pink, *Nat* expression is shown in opaque black, the **a**-specific genes *STE2*, *AXL1*, *ASG7*, *BAR1*, *STE6* and *MAT**a**2* are shown in green, and all other mRNAs are shown in translucent black. Replicate 1 vs. Replicate 1 plot is the same as shown in main text Fig. 2d, all other plot are pairwise comparisons of the other biological replicates performed in parallel in this experiment.

b. mRNA-seq of wildtype *W. anomalus* α cells (*MATα*) vs. second isolate of α cells with *MATα2* deleted (*MATα2 matα2-Δ2)*. *MATα2* expression in transcripts per million (tpm) is shown in pink, *Nat* expression is shown in opaque black, the **a**-specific genes *STE2*, *AXL1*, *ASG7*, *BAR1*, *STE6,* and *MAT**a**2* are shown in green, and all other mRNAs are shown in translucent black. The wildtype *α* cell data is the same as in Supplementary Fig. 3a.

c. Expression levels of genes of particular interest from mRNA-seq experiments plotted above, in transcripts per million (tpm). Shown are **a**-specific genes (*STE2*, *AXL1*, *ASG7*, *BAR1*, and *STE6),* mating-type transcriptional regulators (*MAT**a**2, MATα2, and MATα1)*, the drug-resistance marker used to delete *MATα2* (Nat), and the transcript "fgenesh1_pg.2_#_554," which, besides *MATα2, and MATα1* and Nat, is the only significantly differentially regulated transcript shared between both isolates of *MATα2 matα2-Δ vs. MATα2* (at a false discovery rate $q<0.05$, calculated using the RNA-seq analysis software sleuth).

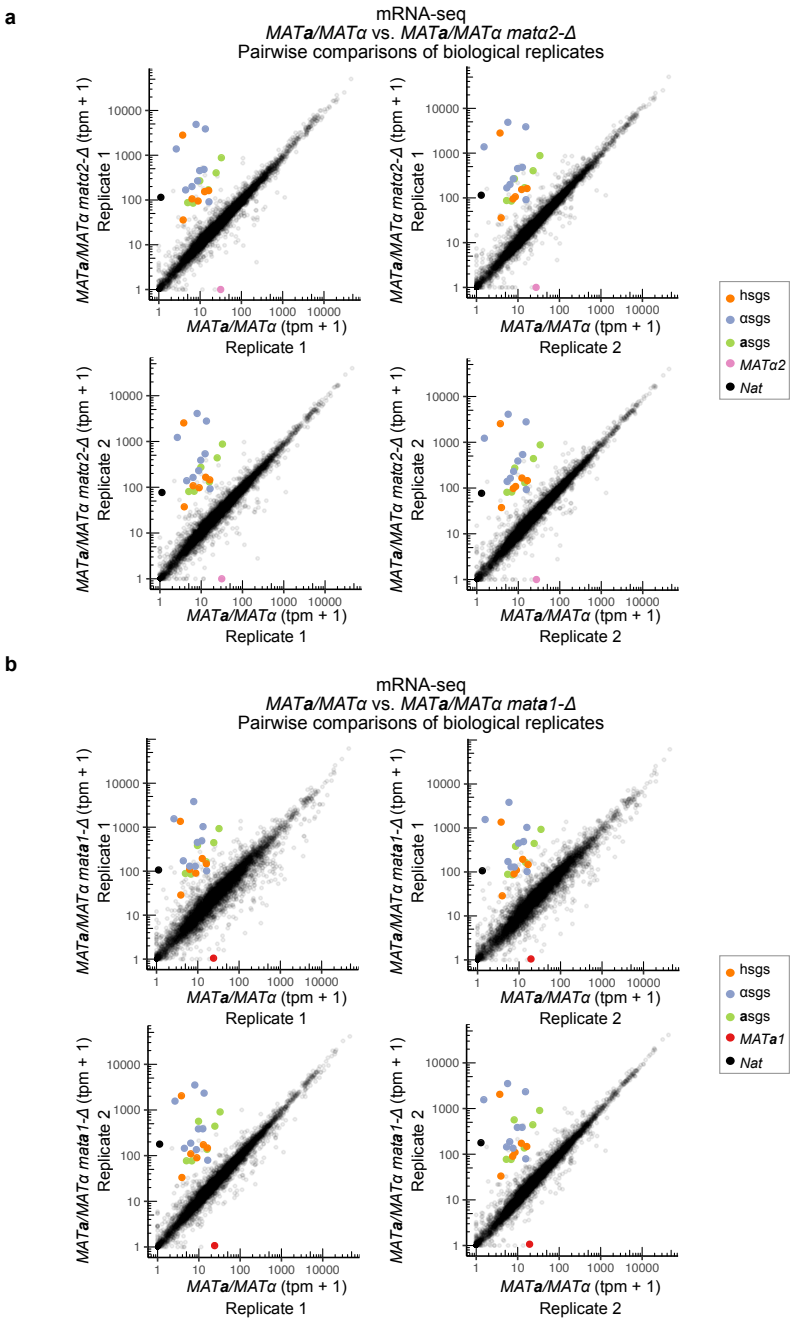


Supplementary Figure 4. *W. anomalus* a-specific genes are activated by Matα2, the ancestral mode of gene regulation

a. We identified matches to both the *C. albicans* Mat**a**2/Mcm1 and the *S. cerevisiae* Matα2/Mcm1 binding sites in the promoters of the *W. anomalus, W. ciferrii,* and *C. jadinii* **a**-specific genes (Bailey and Gribskov, 1998). These putative *cis*-regulatory sequences were cloned into the *S. cerevisiae* P_{CYC1}-GFP reporter and transformed into *S. cerevisiae* **a** and α cells to test for Matα2-dependent transcriptional repression. While the *S. cerevisiae STE2 cis*-regulatory sequence supports transcriptional repression by Matα2, none of the *W. anomalus* clade **a**-specific gene *cis*-regulatory sequences supported repression by Matα2 to the same extent. Only one of the 14 *cis*-regulatory sequences tested, found upstream of a paralog of the *W. anomalus* **a**-mating pheromone gene *MFA* ("*MFA8*"), showed a

statistically significant difference between **a** and α cells with 1.75-fold repression ($p < 0.05$). This sequence is also the best match to the *S. cerevisiae* Matα2/Mcm1 binding site that we identified in the *W. anomalus* clade. Given the overall evidence for a lack of *W. anomalus* clade Matα2 repression of the **a**-specific genes, we do not believe that this single instance of partial repression by *S. cerevisiae* Matα2 reflects actual or vestigial Matα2/Mcm1 repression in *W. anomalus*.

b. *De novo* motifs for the *W. anomalus* clade **a**-specific genes were generated using the motif-elicitation software MEME, and compared to both the *C. albicans* **a**-specific gene *cis*-regulatory sequence (Mat**a**2-Mcm1 binding site) and the *S. cerevisiae* **a**-specific gene *cis*-regulatory sequence (Matα2-Mcm1 binding site) using Tomtom, a tool within the MEME suite (Bailey et al., 2009; Gupta et al., 2007). While the *W. anomalus* clade **a**-specific gene *cis*-regulatory sequence shares homology with both the *S. cerevisiae* and *C. albicans* sequences due to the shared Mcm1-binding site, it is more similar to the *C. albicans* site than the *S. cerevisiae* one.

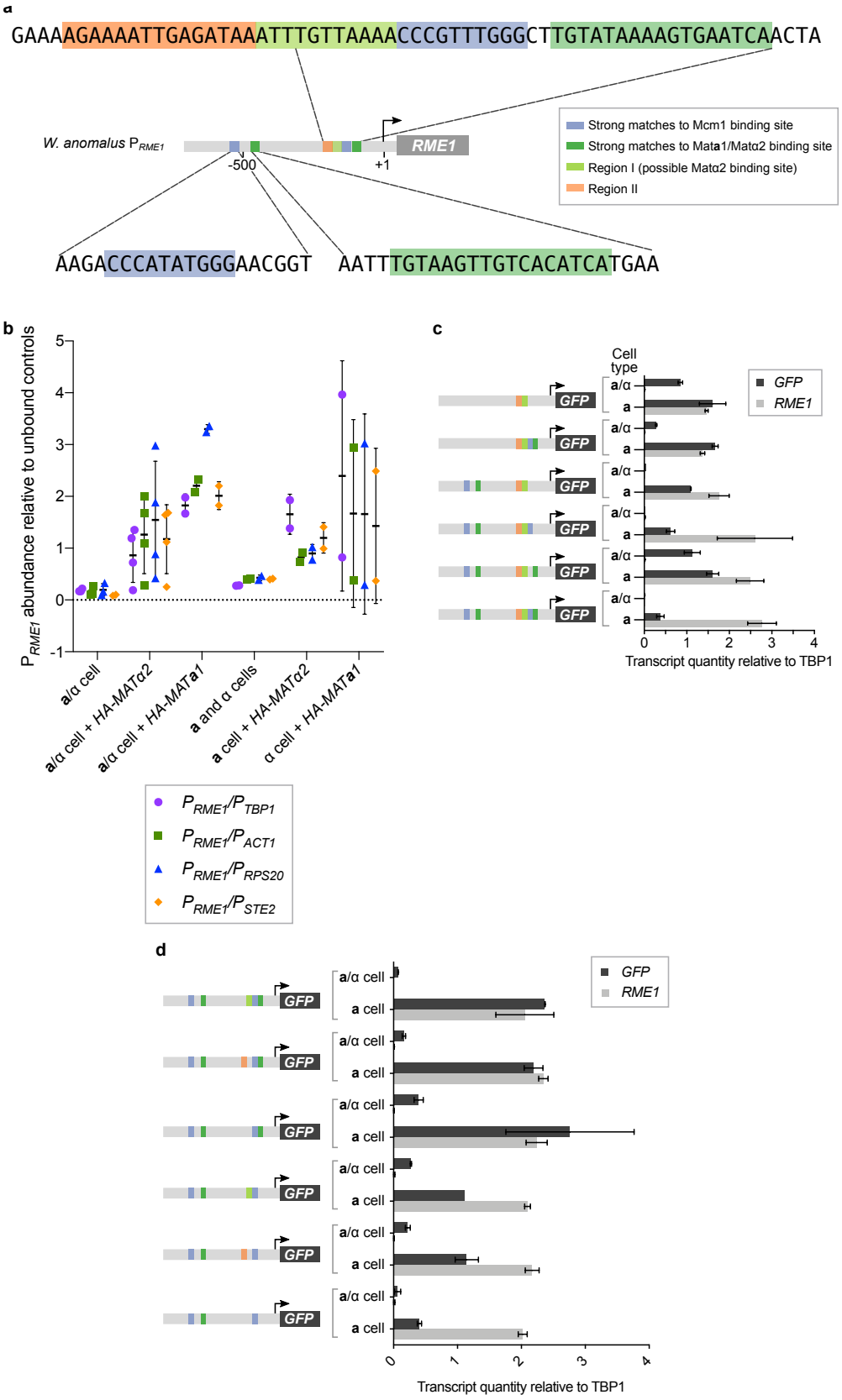


Supplementary Figure 5. The *W. anomalus* haploid-specific genes are repressed by Matα2 and Mata1

a. mRNA-seq of wildtype *W. anomalus* **a**/α cells (*MAT**a**/MATα*) vs. **a**/α cells with *MATα2* deleted (*MAT**a**/MATα matα2-Δ)*. *MATα2* expression in transcripts per million (tpm) is shown in pink, the **a**-specific genes *STE2*, *AXL1*, *ASG7*, *BAR1*, *STE6* and *MAT**a**2* are shown in green, the haploid-specific genes *STE4*, *GPA1*, *FUS3*, *SST2*, *RME1*, and *FAR1* are shown in orange, the α-specific genes *SAG1*, *STE3*, *STE13*, *MFα* (multiple copies), *AFB1* (multiple copies), and *MATα1* are shown in blue-gray, the selective marker used to delete *MATα2* is shown in opaque black, and all other mRNAs are shown in translucent black. The **a**-, α- and haploid-specific genes are more highly expressed when *MATα2* is

deleted. Replicate 1 vs. Replicate 1 plot is the same as shown in main text Fig. 3a, all other plot are pairwise comparisons of the other biological replicates performed in parallel in this experiment.

b. mRNA-seq of wildtype *W. anomalus* **a**/α cells (*MAT**a**/MATα*) vs. **a**/α cells with *MAT**a1*** deleted (*MAT**a**/MATα mat**a**1-Δ)*. Genes are highlighted as above, except *MAT**a1*** is shown in red and *MATα2* is not highlighted. As when *MATα2* is deleted from an **a**/α cell, the **a**-, α- and haploid-specific genes are more highly expressed when *MAT**a1*** is deleted. Replicate 1 vs. Replicate 1 plot is the same as shown in main text Fig. 3b, all other plot are pairwise comparisons of the other biological replicates performed in parallel in this experiment.


a
GAAAAGAAAATTGAGATAAATTTGTTAAAACCCGTTTGGGCTTGTATAAAAGTGAATCAACTA
W. anomalus P RME1
RME1
-500
+1
Strong matches to Mcm1 binding site
Strong matches to Mata1/Matα2 binding site
Region I (possible Matα2 binding site)
Region II
AAGACCCATATGGGAACGGT
AATTTGTAAGTTGTCACATCATGAA
b
P RME1 abundance relative to unbound controls
a/α cell
a/α cell + HA-MATα2
a/α cell + HA-MATa1
a and α cells
a cell + HA-MATα2
α cell + HA-MATa1
P RME1/P TBP1
P RME1/P ACT1
P RME1/P RPS20
P RME1/P STE2
c
Cell type
GFP
RME1
Transcript quantity relative to TBP1
d
a/α cell
a cell
GFP
RME1
Transcript quantity relative to TBP1

Supplementary Figure 6. Mcm1, Matα2, and Mata**1 are directly required for *RME1* repression in** a/**α cell**

a. Diagram of the sequence upstream of the *RME1* coding sequence indicating the two strongest matches to Mat**a**1/Matα2 (green) and Mcm1 (blue) binding sites detected, as well as their location relative to the transcription start site. In addition, the smaller lime green box (Region I) indicates a possible Matα2 binding site, and Region II indicates another stretch of sequence to the left of the proximal Mcm1 site that we experimented with. The TGT residues within Region I may be a possible Matα2 binding site and are positioned relative to the adjacent Mcm1 binding site as they would be in a *S. cerevisiae* Matα2/Mcm1 **a**-specific gene *cis*-regulatory site.

b. Matα2 and Mat**a**1 were N-terminally HA-tagged and transformed into *W. anomalus* cells. Chromatin immunoprecipitation (ChIP) using anti-HA antibodies was performed with wildtype *W. anomalus* cells as an untagged control, followed by qPCR to measure abundance of P_{RME1} relative to several unbound regions as controls (P_{TBP1} in purple circles, P_{ACT1} in green squares, P_{RPS20} in blue triangles, and P_{STE2} in orange diamonds). Results from individual ChIP samples are plotted as dots, and lines show the mean and SD for each strain background.

c. Expression levels of endogenous *RME1* transcript and P_{RME1}-GFP reporter (*GFP*) in *W. anomalus* **a** and **a**/α cells measured by RT-qPCR. Mat**a**1-Matα2 (green) and/or Mcm1 (blue) binding sites were deleted from P_{RME1}-GFP individually and in combination, and the reporters were transformed into an **a** cell. Orange and lime green boxes show the Region I and II *cis*-sequences to the left of the proximal Mcm1 start site that are unaltered in this experiment. The **a** cells were crossed to α cells to form **a**/α cells, and expression levels of the reporter and the endogenous transcript were measured in both cell types. The proximal Mcm1-binding site with respect to the transcription start site is necessary for *RME1* repression in the **a**/α cell. Quantities are mean and SD of two cultures grown and measured in parallel, normalized to expression of the housekeeping gene *TBP1*. This is a repeat of the experiment shown in Fig. 3e, with independent genetic isolates of the strains.

d. Expression levels of endogenous *RME1* transcript and P_{RME1}-GFP reporter (*GFP*) in *W. anomalus* **a** and **a**/α cells measured by RT-qPCR. Possible Mat**a**1-Matα2 binding sites (orange and lime green boxes) immediately to the left the proximal Mcm1 site were deleted from P_{RME1}-GFP individually and in combination, and the reporters were transformed into an **a** cell. The **a** cells were then crossed to α cells to form **a**/α cells, and expression levels of the reporter and the endogenous transcript were measured in both cell types. While the sequences to the right of the proximal Mcm1 site strongly resemble the Mat**a**1-Matα2 binding site from other yeast species, it is actually the sequences to the left of this Mcm1 *cis*-regulatory site that are necessary for complete *RME1* repression in the **a**/α cell (Booth et al., 2010). Quantities are mean and SD of two cultures grown and measured in parallel, normalized to expression of the housekeeping gene *TBP1*.

Materials and Methods

Growth conditions and media

All strains were grown on yeast extract peptone dextrose media (YEPD) at 30°C unless otherwise noted.

Identification of the Phaffomycetaceae as candidate species

Four yeast species with sequenced genomes form a monophyletic group that branches outside of the last common ancestor of the Saccharomycetaceae (the group that contains the *Saccharomyces* and *Kluyveromyces* species, also known as the "*S. cerevisiae* clade"), as well as the Saccharomycodaceae (an outgroup to the Saccharomycetaceae that includes *Hanseniaspora valbyensis* and other *Hanseniaspora* species). At the time this study was completed this clade, termed the Phaffomycetaceae, consisted of two pairs of sister species: *W. anomalus* and *W. ciferrii*, and *C. jadinii* and *C. fabianii* (Riley et al., 2016; Shen et al., 2016). Multiple phylogenomic analyses have placed the Phaffomycetaceae as branching outside of the Saccharomycetaceae and the Saccharomycodaceae with high confidence (Riley et al., 2016; Shen et al., 2016). Other Phaffomycetaceae species that lacked sequenced genomes at the time this study was conducted were excluded from this study. In this study we use the terms "*W. anomalus* clade" and Phaffomycetaceae interchangeably.

Identification of Phaffomycetaceae mating type transcriptional regulators

The genome sequenced *W. anomalus* strain (NRRL Y-366-8) is an **a** cell. The **a**-mating type locus (*MAT**a***), which encodes the **a**-mating type transcriptional regulators *MAT**a**1* and *MAT**a**2*, is located between the *DIC1* and *SLA2* genes, which are syntenic with the mating type locus in many Saccharomycotina species (Riley et al., 2016; Wolfe and Butler, 2017). Unlike *S. cerevisiae*, *W. anomalus* is not known to undergo mating-type switching, and there is no evidence for a silenced mating-type locus or HO endonuclease in its genome (Kurtzman and Fell, 1998;

Wolfe and Butler, 2017). To determine the sequence of the α-mating type locus (*MATα*), we designed oligonucleotides to PCR amplify the region between the *DIC1* and *SLA2* coding sequences of NRRL Y-2153-4, another *W. anomalus* isolate that has been reported to mate with the sequenced **a** cell. This sequencing revealed the genes for the α-mating type transcriptional regulators *MATα1* and *MATα2*. All of the *W. anomalus* mating transcriptional regulator genes except for *MATα1* contain introns, so we amplified and sequenced *MATα2*, *MAT**a**1* and *MAT**a**2* from cDNA to determine their exon coding sequences. Orthology of the *MAT* genes was further confirmed by best reciprocal BLAST to other Saccharomycotina sequences.

The genome sequenced strain of *W. ciferrii* (NRRL Y-1031) is also an **a** cell, so we used the same approach to determine the sequence of the *W. ciferrii MATα* locus. We amplified and sequenced the region between *W. ciferrii DIC1* and *SLA2* in the complementary mating types NRRL# Y-1031-11 and NRRL# Y-1031-27, and found them to be an **a** cell and an α cell, respectively. The *C. fabianii* genome sequenced strain, NRRL Y-1871, is an α cell, so we were able to directly determine the sequence of *MATα* from the existing sequenced genome. The sequenced *C. jadinii* isolate is homothallic and tetraploid, and the genome therefore encodes both *MAT**a*** and *MATα*. All four Phaffomycetaceae *MATα2* genes contain introns, so for each gene we sequenced the cDNA transcript in order to determine its exonic coding sequence.

Identification of cell-type specific genes

Phaffomycetaceae haploid, **a**, and *α*-specific genes were identified by best reciprocal TBLASTN to other Saccharomycotina cell-type specific genes, and with the Yeast Genome Analysis Pipeline (YGAP) (Altschul et al., 1990; Proux-Wéra et al., 2012). In *W. anomalus* the expression patterns of the cell-type specific genes were confirmed by measuring transcript abundance using the NanoString nCounter sytem (method described below) (Supplementary Fig. 2).

Strain construction in *W. anomalus*

The genome-sequenced strain of *W. anomalus*, NRRL Y-366-8, is also the *W. anomalus* type strain and an **a** cell. To generate an **a**/α strain of *W. anomalus*, NRRL Y-366-8 was plated on media containing 5-FOA to select for spontaneous uracil auxotrophs. An **a** cell auxotrophic for uracil was then crossed to NRRL Y-2153-4, an α cell with a naturally occurring arginine auxotrophy. The resulting **a**/α cell (yCSB 8) was selected for by its ability to grow on synthetic defined media with both arginine and uracil dropped out. The 2N diploid ploidy of yCSB 8 was confirmed by staining with SYTO™ 13 Green Fluorescent Nucleic Acid Stain followed by flow cytometry, and the presence of both *MAT**a*** and *MATα* mating type loci in yCSB 8 was confirmed by PCR.

*W. anomalus MAT**a*** gene deletions were made in NRRL Y-366-8, and MAT**a**/α deletions were made in yCSB 8. We were unable to efficiently transform the NRRL Y-2153-4 α cell strain. In order to generate a more genetically tractable α cell, we sporulated yCSB 8 by growing it on a sterilized carrot slice at room temperature for at least five days (Kurtzman and Fell, 1998). Carrots were peeled, cut into cylindrical wedges, then placed in glass tubes with 1 ml water and autoclaved at 115°C for 15 minutes. 500 µl of cells from liquid culture were collected, washed in water, then resuspended in 200 µl water, layered on top of a sterilized carrot wedge in a six well plate, and placed at room temperature. Spores were visible after five days, and sporulation can be started either from saturated overnight cultures, or mid-log cultures. Spores were then resuspended in 100 µl of 1 mg/ml zymolyase for 11 minutes at 37°C degrees, then dissected. A resulting α cell that proved to be genetically tractable (yCSB 342) was the parent strain for generating the α cell *MATα2* deletions. In addition, the dissected spores included **a** and α cells with arginine auxotrophies that were used as tester strains in quantitative mating assays (described below).

W. anomalus genes were deleted using a nourseothricin-resistance marker and a freeze-thaw transformation protocol optimized for use in *W. ciferrii* by Schorsch *et al*. (Schorsch et al., 2009). To construct deletion cassettes, we cloned homology arms into pCS.ΔLig4, a plasmid generated by Schorsch *et al*. (Schorsch et al., 2009). pCS.ΔLig4 contains *nat*, the nourseothricin-resistance gene, codon-optimized for expression in *W. ciferrii*, and under control of *W. ciferrii* P_{PDA1} and T_{TEF1}. It also contains a region of the *W. ciferrii LIG4* gene flanked by loxP sites, for inserting pCS.ΔLig4 into *LIG4* by ends-in recombination. To use the *W. ciferrii*-optimized *nat* marker to delete genes in *W. anomalus* we excised the *LIG4* fragment from pCS.ΔLig4 and ligated in 5' homology regions in between the SexAI and SalI restriction sites, and 3' homology regions between the SacI and SpeI restriction sites. Homology regions ranged from approximately 700 basepairs to 2000 basepairs in length. Transformed cells were selected on 50, 200, and 400 μg/ml nourseothricin for NRRL Y-366-8, yCSB 342, and yCSB 8, respectively. Deletions were screened and confirmed by PCR testing for the absence of the deleted ORF, as well as the presence of the junctions between the deletion cassette and the surrounding genome. Approximately 1/500 transformed colonies contained gene deletions, because *W. anomalus*, like many other fungi, seems to favor taking up DNA by non-homologous end joining rather than homologous recombination (Schorsch et al., 2009).

pCS.ΔLig4 also served as the backbone for *MATα2* add-in experiments. A full-length wildtype *MATα2* gene was PCR amplified and inserted into pCS.ΔLig4 between the SexAI and SalI restriction sites, creating pCSB 147. This plasmid was subsequently modified to produce the different alleles of *MATα2* tested in Supplementary Fig. 6 by digesting at either the PstI/XbaI restriction sites or the StyI/BsrGI restriction sites within the *MATα2* coding sequence and ligating in gBlock gene fragments (Integrated DNA Technologies) with specific mutations. These constructs were then linearized by PvuI digestion, transformed into yCSB 5, and allowed to

randomly integrate into the genome. At least two independent integrants of each allele was assayed in any experiment involving a randomly integrated construct in *W. anomalus* in order to control for variability due to integration at different loci.

pCSB 147 was modified for chromatin immunoprecipitation of Matα2 by using a gBlock to introduce a 3x HA tag at the N-terminus of *MATα2* at the PstI/XbaI restriction sites. Similarly, to tag *MAT**a**1* we inserted a full length *MAT**a**1* construct with its promoter and terminator into pCS.ΔLig4 at the SexAI/SalI restriction sites. We then introduced a 3x HA tag at the N-terminus of *MAT**a**1* by inserting a gBlock between the XbaI/NcoI restriction sites. Tagged *HA-MATα2* and *HA-MAT**a**1* constructs were inserted into *W. anomalus* by random integration.

We also used pCS.ΔLig4 as the basis for a hygromycin-resistance marker optimized for expression in *W. anomalus*. *W. ciferrii* P_{PDA1} and T_{TEF1} were amplified from pCS.ΔLig4 and cloned into pUC19 at the BamHI/XbaI and PstI/SphI restriction sites, respectively, to create pCSB 137. The coding sequence of hph (a hygromycin-resistance gene) was then codon-optimized for expression in *W. anomalus* using a codon usage table generated by the Kazusa DNA Research Institute (http://www.kazusa.or.jp/codon/), synthesized as a gBlock (IDT), and cloned into pCSB 137 at the XbaI/PstI restriction sites to create pCSB 182. pCSB 182 was used as the backbone for all P_{RME1}-GFP experiments. First, a GFP coding sequence was also codon-optimized for expression in *W. anomalus*, synthesized as a gBlock with the *W. anomalus ACT1* terminator, and cloned into pCSB 182 at the BamHI/SalI restriction sites. 944 basepairs of the *RME1* regulatory region (immediately upstream of the beginning of the *RME1* ORF) were cloned upstream of GFP by Gibson assembly. Modifications of P_{RME1} were subsequently made by inserting gBlocks between the DraIII/SpeI restriction sites that naturally occurred within P_{RME1}. All P_{RME1}-GFP constructs were linearized by digestion with SphI and allowed to randomly integrate into the genome of a *W.*

anomalus **a** cell (NRRL Y-366-8). Transformants were selected on YEPD plates with 300 μg/ml hygromycin.

Strain construction in *S. cerevisiae*

All *S. cerevisiae* experiments were performed in the W303 strain, and transformations were performed by a published lithium acetate/single-stranded carrier DNA/polyethylene glycol method (Gietz and Woods, 2002). *MATα2* was deleted from the W303 α cell using a KanMX marker amplified with homology to *MATα2*.

MATα2 expression constructs for **a**-specific gene transcriptional repression assays were cloned into pNH 604, a plasmid that integrates into the *S. cerevisiae* genome in single copy at the *trp1* locus using a *TRP1* marker (Baker et al., 2012; Youk and Lim, 2014). pNH 604 was digested with SacII and KpnI to excise the Tef overexpression promoter. *MATα2* constructs were amplified from cDNA or synthesized as gBlock synthetic gene fragments (Integrated DNA Technologies) and fused to the *S. cerevisiae MATα2* promoter by fusion PCR, then inserted into pNH 604 between the SacII/KpnI restriction sites. pNH 604 were linearized by PmeI digestion, transformed into the *MATα matα2-Δ* strain, and selected for on SD-trp plates.

The P_{CYC1}-GFP reporter was designed to integrate in single copy at the *ura3* locus using the hph hygromycin resistance marker. It was adapted from the GFP reporter pTS 61 and the beta-galactosidase reporter pLG699z (Guarente and Ptashne, 1981; Sorrells et al., 2015). P_{CYC1} was inserted upstream of GFP, with the addition of a KpnI restriction site, allowing *cis*-regulatory elements to be inserted between the XhoI and KpnI sites in P_{CYC1}, between the upstream activating sequence and transcription start site of the promoter. The hph marker is up stream of P_{CYC1}-GFP, and the two together are flanked by homology to *URA3*. Overlapping oligos with the **a**-specific gene *cis*-regulatory elements and phosphorylated sticky ends were designed, annealed, and ligated into the XhoI/KpnI sites.

Design of *MATα2* mutant alleles

To construct chimeric alleles of *MATα2*, we defined the five regions of the Matα2 protein based on the *S. cerevisiae* sequence as previously described (region 1: amino acid 1-21, region 2: amino acid 22-108, region 3: amino acid 109-127, region 4: amino acid 128-188, region 5: amino acid 189-210) (Baker et al., 2012). To introduce a loss of function mutation into region 5 of *W. anomalus MATα2*, we introduced a L215A point mutation (Li et al., 1995)

Gene expression profiling by mRNA-seq

For gene expression profiling by RNA-seq, single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in fresh YEPD, and then grown for 4-4.5 hours at 30°C. Cells were pelleted by centrifugation at 3000 rpm for 10 minutes, then flash frozen in liquid nitrogen and stored at -80°C. RNA was extracted, purified, and DNase-treated using an Ambion RiboPure RNA Purification kit for yeast. Quality of total RNA was evaluated on an Agilent Bioanalyzer using an Agilent RNA 6000 Pico Kit. Two rounds of polyA selection were performed using the Qiagen Oligotex mRNA Mini Kit, and mRNA quality was evaluated on an Agilent Bioanalyzer using an Agilent RNA 6000 Pico Kit. mRNA samples were then concentrated using a Zymo RNA Clean and Concentrator Kit. cDNA synthesis and library preparation were performed using a NEBNext Directional RNA-seq kit. Libraries were sequenced using single end 50 basepair reads on the Illumina HiSeq 4000 sequencing system in the UCSF Center for Advanced Technologies. Reads were pseudoaligned to all *W. anomalus* coding sequences using kallisto, and differential expression analyses were performed using sleuth (Bray et al., 2016; Pimentel et al., 2017). Coding sequences were downloaded from the Joint Genome Institute annotation of genes for *Wickerhamomyces anomalus* NRRL Y-366-8 v1.0 (Riley et al., 2016). For differential expression analyses, we focused on the expression patterns of candidate haploid, **a**, and α-specific

genes, except in the case of *matα2-Δ* in α cells, where we used a false discovery rate cut off of q < 0.05 to search for differentially expressed genes.

Gene expression profiling with NanoString nCounter

To measure transcript abundance with the NanoString nCounter system (www.nanostring.com), single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in fresh YEPD, and then grown for 4-4.5 hours at 30°C. Cells were pelleted by centrifugation for 10 minutes at 3000 rpm, then flash frozen in liquid nitrogen and stored at -80°. RNA was extracted, purified, and DNase-treated using an Ambion RiboPure RNA Purification kit for yeast. NanoString quantification was performed by the UCSF Center for Advanced Technologies using a set of probes designed by NanoString to target the *W. anomalus* mating-type transcriptional regulators, haploid-specific genes, **a**-specific genes, α-specific genes, and a set of non-differentially expressed reference genes (*GAPDH, TBP1, SPC98 and MRPL49*). The NanoString probe for the **a**-mating pheromone gene (*MFA*) targets at least seven paralogs of this gene in *W. anomalus*, because their transcripts are too small and similar in sequence to distinguish from each other by this method. Transcript counts were normalized across samples to the quantities of the non-differentially expressed reference genes.

GFP reporter assays

Single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in SD supplemented with amino acids and uracil, and then grown for 4-4.5 hours at 30°C. GFP fluorescence was measured on a BD LSR II flow cytometer. 10,000 cells were measured per sample in each experiment. Cells were gated to exclude debris, and the mean GFP fluorescence was calculated for each reporter strain sample. Three independent isolates of each reporter strain were measured.

Quantitative mating assays

Quantitative mating assays were performed based on methods described for other yeast species (Miller and Johnson, 2002). Strains of interest (for example, gene deletions and their parent strains) were plated on 5-FOA to select for uracil auxotrophs (as described above), and tester strains of the opposite mating type with arginine were isolated from sporulating the **a**/α strain yCSB 8. Single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.15 in fresh YEPD, and then grown to mid-log phase and combined in a final volume of 5 ml. 1 ml (about $1x10^7$ cells) of the mating type present in excess was combined with 200 µl of the strain to be measured, for a final ratio of 5:1 strain in excess to limiting strain, by volume. The mating mixes were then deposited onto 0.8 µm nitrocellulose filters using a Millipore 1225 Vacuum Sampling Manifold. The filters were then placed on a YEPD agar plate and incubated for 5 hours at 30°C. The cells were resuspended by placing the filters in 5 ml water, and cells dispersed by vortexing. Dilutions were plated on SD-arg-ura plates to select for conjugants, and SD-ura to select for the limiting parent as well as to conjugants. After 2-3 days of growth at 30°C, colonies were counted, and the mating efficiency was calculated by the following formula: Mating efficiency = (mating products) / (mating products + limiting tested parents) x 100%. In other words, Mating efficiency = (number of colonies on SD-arg-ura)/(number of colonies on SD-ura) x 100%.

***Cis*-regulatory motif discovery and searching**

A *cis*-regulatory motif for the Phaffomycetaceae **a**-specific genes was generated using MEME (Bailey et al., 2009). We used sequences 500 basepairs upstream of the **a**-specific gene orthologs in *W. anomalus* and *W. ciferrii*, and used MEME to search for motifs between 24 and 27 bases wide occurring zero or one time in those sequences. These sequences were from the *W. ciferrii STE2*, *STE6*, and *AXL1* genes, and the *W. anomalus STE2*, *STE6*, *ASG7*, *AXL1*, *BAR1*, and *STE14* genes. This returned a motif that resembled a Mat**a**2-Mcm1 binding site. We used the

TOMTOM tool in the MEME suite to align and compare this *W. anomalus*/*W. ciferrii* **a**-specific gene motif to both the *S. cerevisiae* Matα2-Mcm1 motif and the *C. albicans* Mat**a**2/Mcm1 motif (Gupta et al., 2007). We searched for Matα2, Mat**a**1, Mat**a**2, and Mcm1 *cis*-regulatory motifs using the MAST and FIMO tools in the MEME suite.

RT-qPCR

Single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in fresh YEPD, then grown for 4-4.5 hours at 30°C. Cells were pelleted by centrifugation for 10 minutes at 3000 rpm, then flash frozen in liquid nitrogen and stored at -80°. RNA was purified using the MasterPure™ Yeast RNA Purification Kit from Epicentre. cDNA was synthesized with random hexamer primers using SuperScript® IV reverse transcriptase. qPCR was performed using BioRad SYBR Green Master Mix on a BioRad CFX Connect™ Real-Time PCR Detection System. Each sample was quantified in three technical replicates, and each experiment was performed with two biological replicates per sample. Relative quantifications were determined using a standard curve, and quantifications for *RME1* and *GFP* were normalized to the non-differentially regulated housekeeping gene *TBP1*.

Chromatin-immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed as previously described (Nocedal et al., 2017; Sorrells et al., 2015). Briefly, single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in fresh YEPD, then grown at 30°C to OD_{600} = 0.4. Cells were then crosslinked with formaldehyde, quenched with glycine, washed with TBS (20mMTris/HCl pH 7.4, 150 mM NaCl), flash frozen in liquid nitrogen, and stored at -80°C. Cells were lysed at 4°C in lysis buffer (50 mM HEPES/KOH pH 7.5, 140 mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-Deoxycholate) plus protease inhibitors (Roche Complete Protease Inhibitor Cocktail EDTA-free) with 0.5 mm glass beads on a vortex for two hours. Recovered chromatin was then sheared

via sonication on a Diagenode Biorupter (three 10 minute periods of sonication on level 5 for 30 seconds, then off for 1 minute). 10 mg of anti-HA 12CA5 from Roche was used to bind the HA-tagged transcription factors overnight at 4°C in lysis buffer plus protease inhibitors, and protein-G sepharose beads were used to collect immunoprecipitated DNA. Immunoprecipitated DNA was then recovered by washing with lysis buffer, wash buffer (10mM Tris/HCl pH 8.0, 250mM LiCl, 0.5% NP-40, 0.5% Na-Deoxycholate, 1mM EDTA), and TE, then collected in elution buffer (50mM Tris/HCl pH8.0, 10mM EDTA, 1%SDS). Crosslinking was reversed by Proteinase K treatment for two hours at 37°C, then16 hours at 65°C. DNA was then cleaned up using the Qiagen MinElute kit.

*HA-MAT**a**1* and *HA-MATα2* constructs were inserted into the *W. anomalus* genome via random integration. Two to four independent isolates of each strain were processed, and two samples of untagged **a**/α cells, and one each of untagged **a** and untagged α cells were also processed as negative controls. Abundance of the *RME1* promoter region in tagged vs. untagged samples was measured by qPCR compared to the promoter regions of the housekeeping genes *TBP1*, *ACT1*, and *RPS20*, and the **a**-specific gene *STE2*.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Bailey, T.L., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 48–54.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

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

Baker, C.R., Tuch, B.B., and Johnson, A.D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc. Natl. Acad. Sci. U.S.a. *108*, 7493–7498.

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

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

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

Dalal, C.K., Zuleta, I.A., Mitchell, K.F., Andes, D.R., El-Samad, H., and Johnson, A.D. (2016). Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. Elife *5*, e18981.

Geiss, G.K., Bumgarner, R.E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D.L., Fell, H.P., Ferree, S., George, R.D., Grogan, T., et al. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat. Biotechnol. *26*, 317–325.

Gietz, R.D., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Meth. Enzymol. *350*, 87–96.

Gray, M.W., Lukes, J., Archibald, J.M., Keeling, P.J., and Doolittle, W.F. (2010). Irremediable Complexity? Science *330*, 920–921.

Guarente, L., and Ptashne, M. (1981). Fusion of Escherichia coli lacZ to the cytochrome c gene of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.a. *78*, 2199–2203.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

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

Ho, C.-Y., Adamson, J.G., Hodges, R.S., and Smith, M. (1994). Heterodimerization of the yeast MATa1 and MAT alpha 2 proteins is mediated by two leucine zipper-like coiled-coil motifs. The EMBO Journal *13*, 1403–1413.

KOMACHI, K., Redd, M.J., and Johnson, A.D. (1994). The Wd Repeats of Tup1 Interact with the Homeo Domain Protein Alpha-2. Genes Dev. *8*, 2857–2867.

Komachi, K. (1996). The Interaction Between Alpha 2 and Tup1.

Kurtzman, C., and Fell, J.W. (1998). The Yeasts - A Taxonomic Study (Elsevier).

Kurtzman, C.P. (2011). Phylogeny of the ascomycetous yeasts and the renaming of Pichia anomala to Wickerhamomyces anomalus. Antonie Van Leeuwenhoek *99*, 13–23.

Lavoie, H., Hogues, H., Mallick, J., Sellam, A., Nantel, A., and Whiteway, M. (2010). Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol *8*, e1000329.

Li, T., Jin, Y., Vershon, A.K., and Wolberger, C. (1998). Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. Nucleic Acids Res. *26*, 5707–5718.

Li, T., Stark, M.R., Johnson, A.D., and Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. Science *270*, 262–269.

Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F., and Gray, M.W. (2011). How a neutral evolutionary ratchet can build cellular complexity. IUBMB Life *63*, 528–537.

Lynch, V.J., and Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. Evolution *62*, 2131–2154.

Mak, A., and Johnson, A.D. (1993). The carboxy-terminal tail of the homeo domain protein alpha 2 is required for function with a second homeo domain protein. Genes Dev. *7*, 1862–1870.

Mead, J., Zhong, H.L., Acton, T.B., and Vershon, A.K. (1996). The yeast alpha 2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. Mol. Cell. Biol. *16*, 2135–2143.

Mead, J., Bruning, A.R., Gill, M.K., Steiner, A.M., Acton, T.B., and Vershon, A.K. (2002). Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. Mol. Cell. Biol. *22*, 4607–4621.

Miller, M.G., and Johnson, A.D. (2002). White-Opaque Switching in Candida albicans Is Controlled by Mating-Type Locus Homeodomain Proteins and Allows Efficient Mating. Cell *110*, 293–302.

Nocedal, I., Mancera, E., and Johnson, A.D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. Elife *6*, 1.

PHILLIPS, C.L., Stark, M.R., Johnson, A.D., and DAHLQUIST, F.W. (1994). Heterodimerization of the Yeast Homeodomain Transcriptional Regulators Alpha-2 and A1 Induces an Interfacial Helix in Alpha-2. Biochemistry *33*, 9294–9302.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690.

Proux-Wéra, E., Armisén, D., Byrne, K.P., and Wolfe, K.H. (2012). A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. BMC Bioinformatics *13*, 237.

Riley, R., Haridas, S., Wolfe, K.H., Lopes, M.R., Hittinger, C.T., Goeker, M., Salamov, A.A., Wisecaver, J.H., Long, T.M., Calvey, C.H., et al. (2016). Comparative genomics of biotechnologically important yeasts. Proc. Natl. Acad. Sci. U.S.a. *113*, 9882–9887.

Sauer, R.T., Smith, D.L., and Johnson, A.D. (1988). Flexibility of the yeast alpha 2 repressor enables it to occupy the ends of its operator, leaving the center free. Genes Dev. *2*, 807–816.

Sayou, C., Monniaux, M., Nanao, M.H., and Moyroud, E. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science.

Schorsch, C., Köhler, T., and Boles, E. (2009). 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.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C.P., Hittinger, C.T., and Rokas, A. (2016). Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. G3 (Bethesda) *6*, 3927–3939.

Sorrells, T.R., and Johnson, A.D. (2015). Making sense of transcription networks. Cell *161*, 714–723.

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

Stern, D.L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? Evolution *62*, 2155–2177.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J. Mol. Evol. *49*, 169–181.

Tan, S., and Richmond, T.J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature *391*, 660–666.

Tsong, A.E., Miller, M.G., Raisner, R.M., and Johnson, A.D. (2003a). Evolution of a Combinatorial Transcriptional Circuit. Cell *115*, 389–399.

Tsong, A.E., Miller, M.G., Raisner, R.M., and Johnson, A.D. (2003b). Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell *115*, 389–399.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative transcriptional circuits with identical logic. Nature *443*, 415–420.

Tuch, B.B., Galgoczy, D.J., Hernday, A.D., Li, H., and Johnson, A.D. (2008). The evolution of combinatorial gene regulation in fungi. PLoS Biol *6*, e38.

Wagner, A. (2005). Robustness, evolvability, and neutrality. FEBS Lett. *579*, 1772–1778.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet *13*, 59–69.

Wolfe, K.H., and Butler, G. (2017). Evolution of Mating in the Saccharomycotina. Annu. Rev. Microbiol. *71*, 197–214.

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

Youk, H., and Lim, W.A. (2014). Secreting and Sensing the Same Molecule Allows Cells to Achieve Versatile Social Behaviors. Science *343*, 1242782–1242782.

Chapter Three

Multiple regions of the *Wickerhamomyces anomalus* Matα2 DNA-binding domain contribute to its inability to repress a-specific gene transcription

Introduction

Changes in gene expression patterns are an important mechanism by which organisms evolve. Gene expression patterns are typically controlled by transcriptional circuits made up of sequence-specific DNA-binding proteins that regulate gene expression, and *cis*-regulatory elements adjacent to their direct target genes. Given substantially longer evolutionary timespans, the coding sequences of sequence specific DNA binding proteins themselves can evolve, which can have profound consequences for the regulation of whole suites of genes at once. Previously, I showed that the yeast mating-type transcriptional regulator Matα2 underwent an evolutionary change in which it maintained its ancient evolutionary function (repression of the haploid-specific genes) while it evolved new regions capable of forming specific protein-protein interactions with other transcriptional regulatory proteins (Chapter 2). Subsequently, this new assemblage of Matα2 with other transcriptional regulatory proteins was co-opted for the regulation of a new set of genes (the **a**-specific genes). I inferred this evolutionary pathway by experimentally characterizing Matα2 in *Wickerhamomyces anomalus*, a member of the clade of Phaffomycetaceae species that diverged from the model yeast *Saccharomyces cerevisiae* before the point in time when Matα2 evolved to repress the **a**-specific genes.

I showed that in *W. anomalus*, Matα2 does not repress the **a**-specific genes, despite the fact that it has protein-protein interaction regions capable of binding Tup1 and Mcm1, the other transcriptional regulatory proteins necessary for **a**-specific gene repression in *S. cerevisiae*. There are two independent mechanisms that account for *W. anomalus* Matα2's lack of **a**-specific gene repression. First, the **a**-specific gene cis-regulatory elements in *W. anomalus* are Mat**a**2-Mcm1 binding sites, not Matα2-Mcm1 binding sites. We know this is the case because the *W. anomalus* **a**-specific genes are activated by Mat**a**2, and because *S. cerevisiae* Matα2, a bona fide **a**-specific

gene repressing-Matα2, cannot repress transcription at *W. anomalus* **a**-specific gene *cis*-regulatory sites. Thus, *W. anomalus* **a**-specific genes lack the cis-regulatory elements necessary for direct repression by Matα2. In addition, there is a separate reason that *W. anomalus* Matα2 does not repress the **a**-specific genes. Even if *W. anomalus* did have Matα2-Mcm1 binding sites at the **a**-specific genes, Matα2 still would not repress their transcription because its DNA-binding domain itself is not competent for **a**-specific gene repression. In particular, when the sequence for the *W. anomalus* Matα2 DNA-binding domain is placed into the *S. cerevisiae* Matα2 sequence it does away with **a**-specific gene repression. This is true only for the DNA-binding domain, as the four other modules of the *W. anomalus* Matα2 sequence can be substituted into the *S. cerevisiae* Matα2 coding sequence with no effect on **a**-specific gene repression. In the reverse experiment, swapping the *S. cerevisiae* Matα2 DNA-binding domain sequence into the *W. anomalus* Matα2 sequence improves *W. anomalus* Matα2's ability to repress an **a**-specific gene reporter. The same was also true for the Matα2 DNA-binding domain from *Cyberlindnera jadinii*, a second Phaffomycetaceae species. Notably, this is in contrast to the *Candida albicans* and *Pichia pastoris* Matα2 DNA-binding domains, which branched before the Tup1 and Mcm1-interaction regions evolved but are nonetheless competent to repress an **a**-specific gene reporter when the *S. cerevisiae* Matα2 Tup1 and Mcm1-interaction regions are introduced into the protein sequence (Figure 1) (Baker et al., 2012).

The most parsimonious explanation for these observations is that there were evolutionary changes to the Matα2 DNA-binding domain sequence that occurred in the lineage leading to *W. anomalus*, after it diverged from its last common ancestor with *S. cerevisiae.* This raises several questions about Matα2's evolution. First, what are the mutations underlying *W. anomalus* Matα2 DNA-binding domain's inability to repress transcription at a Matα2-Mcm1 binding site? Second,

do those mutations have any bearing on *W. anomalus* Matα2 binding DNA at Mat**a**1-Matα2-Mcm1 binding sites, like those at the *RME1* locus? Are the mutations at all dependent on the other evolutionary changes in the branch leading to *W. anomalus*? To begin to answer these questions, I performed a series of experiments to more closely examine the *W. anomalus* Matα2 DNA-binding domain and compare its amino acid sequence and functional capabilities to the *S. cerevisiae* Matα2 ortholog.

Matα2 is a homeodomain transcription factor, an ancient group of sequence-specific DNA-binding proteins found in a large variety of eukaryotes, and are thought to have been present in the last eukaryotic common ancestor (Derelle et al., 2007). Homeodomain proteins, including the homeotic (Hox) genes in animals, often regulate morphogenesis. As such, Matα2 can serve as a model for how the coding sequence of homeodomain proteins change and co-evolve along with co-regulatory proteins and *cis*-regulatory sites and will likely inform our understanding of other types of substantial evolutionary changes in a diversity of eukaryotes. It is likely that any principles found to underlie Matα2's evolution will be broadly applicable to other transcription factors, and other homeodomain proteins in particular.

The homeodomain DNA-binding domain is made up of a bundle of three alpha-helices. DNA-contacting residues are found in the N-terminal arm and the third alpha helix (also known as the recognition helix) (Li et al., 1995; Tan and Richmond, 1998; Wolberger et al., 1991; 2011). The recognition helix contacts the major groove and is thought to be more important source of base-specific contacts, while the N-terminal arm contacts the minor groove. Residues in both the recognition helix and the N-terminal arm of *S. cerevisiae* Matα2 are necessary for complete repression of both the **a**-specific genes (in combination with Mcm1), and the haploid-specific genes (in combination with Mat**a**1), though mutations in both regions have consistently larger effects on **a**-specific gene repression than haploid-specific gene repression (Mathias et al., 2001).

Both DNA-contacting regions may therefore contain mutations that can explain the changes in *W. anomalus* Matα2.

Results

W. anomalus Matα2 DNA-binding domain shares greater degree of conservation with the *S. cerevisiae* Matα2 DNA-binding domain recognition helix than the N-terminal arm

The *S. cerevisiae* Matα2 DNA-binding domain is made up of 60 amino acid residues. I aligned the *S. cerevisiae* and *W. anomalus* Matα2 protein sequences and found that 25/60 residues are identical between the two species (Figure 2). The recognition helix is particularly well conserved, as it tends to be across all homeodomains, with 100% conservation of the residues that form base-specific contacts (in *S. cerevisiae*, these are S50, N51 and R54), as well as the residues that contact the sugar-phosphate backbone (in *S. cerevisiae*, K57, Q44, W48, K55, and R53) (Li et al., 1998; Mathias et al., 2001; Tan and Richmond, 1998). In contrast, the N-terminal arm is only conserved in two residues (R4 and R7), and the *W. anomalus* sequence has a three amino acid insertion in the N-terminal arm region of the domain.

N-terminal arm is greater contributor to *W. anomalus* Matα2 DNA-binding domain's inability to repress a-specific gene transcription

To ask which DNA-contacting regions of the *W. anomalus* Matα2 DNA-binding domain are responsible for its inability to repress an **a**-specific gene reporter, I performed a series of transcriptional reporter experiments in which I exchanged regions of the *W. anomalus* and *S. cerevisiae* Matα2 DNA-binding domain coding sequences. This approach can reveal the residues responsible for the difference between the two orthologs. I inserted these *MATα2* chimeras into a *S. cerevisiae* α cell with *MATα2* deleted and measured their ability to repress a P_{CYC1}-GFP reporter with a Matα2-Mcm1 binding site from the promoter of the *S. cerevisiae* **a**-specific gene *STE2*. I

again found that the *W. anomalus* Matα2 DNA-binding domain did not support **a**-specific gene repression, recapitulating the results in Chapter 2 (Figure 3). The *S. cerevisiae* Matα2 DNA-binding domain improves the ability of the *W. anomalus* Matα2 sequence to repress **a**-specific gene transcription. This held true when I altered the boundary at which I considered the difference between the Mcm1-interation region of Matα2 and the N-terminus of the DNA-binding domain. That is, I constructed chimeras where the *W. anomalus* DNA-binding domain began at the NGC residues, and they recapitulated the initial chimeric swaps (where the *W. anomalus* DNA-binding domain swapped region began at the IPH residues instead).

Next, I individually exchanged the N-terminal arm and recognition helix sequences between the two orthologs. Both the *W. anomalus* N-terminal arm and recognition helix worsened the *S. cerevisiae* Matα2's ability to repress **a**-specific gene transcription, although the N-terminal arm had a much larger effect. In particular, while the wildtype *S. cerevisiae* Matα2 represses ~98% of transcription, *S. cerevisiae* Matα2 with the *W. anomalus* recognition helix represses ~96%, and *S. cerevisiae* Matα2 with the *W. anomalus* N-terminal arm represses ~70%. These are both small, partial effects when compared to the reduction to ~19% repression when the entire *W. anomalus* DNA-binding domain is swapped into the *S. cerevisiae* Matα2 sequence.

In the reciprocal experiment, I substituted the *S. cerevisiae* N-terminal arm and recognition helix sequences into *W. anomalus* Matα2 and found that neither of these regions conferred an improvement in **a**-specific gene repression. In addition, given the striking features of the *W. anomalus* Matα2 homeodomain N-terminal arm—namely, the extra three amino acids relative to *S. cerevisiae* Matα2—I posited that the extra bulk in that region of the protein provided by those three residues may be sterically hindering the *W. anomalus* Matα2 DNA-binding domain from favorably interacting with DNA. To test this, I also constructed alleles in which I deleted either

the IPH or NGC residues from the *W. anomalus* Matα2 DNA-binding domain sequence and found that these too did not improve a-specific gene repression. Finally, I converted the RGHR residues in the *W. anomalus* N-terminal arm to the homologous *S. cerevisiae* sequence of RRRR. Again, this did not have a large effect on *W. anomalus* Matα2's ability to repress the **a**-specific gene reporter, though it did confer a small improvement from 32% to 47% repression.

Discussion

We previously showed that *W. anomalus* Matα2 does not repress the **a**-specific genes despite the fact that it has the Mcm1 and Tup1-interaction regions necessary to do so and that its DNA-binding domain is not capable of supporting **a**-specific gene repression. Here we ask whether certain coding-sequence changes in the *W. anomalus* Matα2 DNA-binding domain are responsible for this inability to repress **a**-specific gene transcription. By constructing a series of chimeric alleles of Matα2, we found that the N-terminal arm of *W. anomalus* Matα2 has the largest effect in disrupting **a**-specific gene repression. Nonetheless, the N-terminal arm is not sufficient to explain the behavior of the entire *W. anomalus* Matα2 DNA-binding domain, which results in a complete inability to repress **a**-specific gene transcription when substituted into the *S. cerevisiae* Matα2 sequence.

One possible explanation for this disconnect between the sizes of the effects of the N-terminal arm and the entire of the DNA-binding domain is that non-DNA contacting residues in *W. anomalus* Matα2 also contribute to its apparent change in DNA-binding specificity. For examples, mutations in amino acid residues in the first and second alpha helices may have altered the overall shape of the *W. anomalus* Matα2 DNA-binding domain so that it is sterically not compatible for binding DNA with Mcm1. Another possibility is that the N-terminal arm and other residues in the DNA-binding domain work epistatically and have a greater-than-additive effect on

W. anomalus Matα2's inability to repress **a**-specific gene transcription when they're together than when they are separate (Starr and Thornton, 2016). While we have not yet exhaustively mapped the causative mutations that led to *W. anomalus* Matα2's inability to repress **a**-specific gene transcription, we can now be certain that it will evolve multiple regions within the DNA-binding domain.

What do the *W. anomalus* Matα2 DNA-binding domain changes mean for its DNA-binding specificity, and its co-evolution with its co-regulatory proteins? In *W. anomalus* Matα2 does not repress the a-specific genes. Indeed, *W. anomalus* branched before direct **a**-specific gene repression by Matα2 had ever evolved. While *W. anomalus* Matα2 has the ability to bind Mcm1, in *W. anomalus* this interaction functions instead in haploid-specific gene repression. In particular, I found that Matα2, Mat**a**1 and Mcm1 are all necessary to directly bind and repress *RME1* transcription in **a**/α cells, and this is likely one of the few, if not the only direct target of Matα2-Mat**a**1 repression (see Chapter 4). I propose that selection for Matα2's DNA-binding specificity may have relaxed in the lineage leading to *W. anomalus* because of this. In this lineage Matα2 directly binds not one but two sequence-specific DNA proteins, and it is possible that this extra DNA-binding specificity provided by the other two proteins makes it so that the requirement for Matα2's DNA-binding specificity is relatively flexible.

Indeed, we know from genetic experiments in *S. cerevisiae* that many of the DNA-contacting residues in the *S. cerevisiae* Matα2 DNA-binding domain are not actually necessary for haploid-specific gene repression with Mat**a**1 (Mathias et al., 2001). Mutating these residues to alanine results in an inability to repress an **a**-specific gene reporter with Mcm1 but has almost no effect on haploid-specific gene repression with Mat**a**1. Perhaps this is consistent with what we observe in *W. anomalus*. *S. cerevisiae* Mat**a**1 always binds with Matα2 as a heterodimer, so that

all Mata1 regulated genes are also Matα2 regulated genes. In contrast, Matα2 binds with Mcm1 in two copies, one on either side of the Mcm1 homodimer, and Matα2 is the sole provider of the important cell-type specificity needed for **a**-specific gene regulation. Unlike Mat**a**1, Mcm1 is a ubiquitously expressed transcriptional regulator that works in combination with several other transcriptional regulator to govern the expression of hundreds of target genes (Tuch et al., 2008a). Thus, the fewer cell-type specific binding partners Matα2 has, the more its DNA-binding specificity is essential. Perhaps, then, in the lineage leading to *W. anomalus* where Matα2 has a obligate sequence-specific and cell-type specific binding partner in Mat**a**1, and an additional sequence specific binding partner in Mcm1, the DNA-binding domain is under very little selection to maintain a high level of specificity. Its protein sequence can drift without compromising its function, thanks to the strength of its co-factors. This may be one example of what has previously been proposed as a general way in which cooperative binding of transcriptional regulator proteins can facilitate transcriptional circuit evolution (Tuch et al., 2008b).

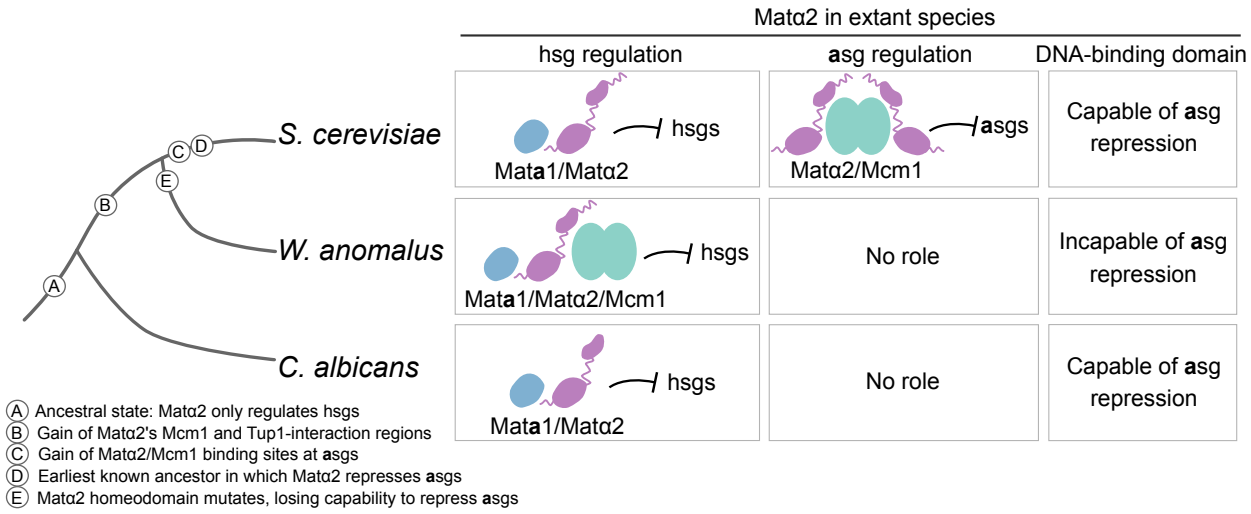


Figure 1. Only *W. anomalus*-clade Matα2 has DNA-binding domains that does not support a-specific gene repression

Matα2's functions in haploid and **a**-specific gene regulation, and DNA-binding traits mapped onto a cladogram of the Saccharomycotina yeast. While *W. anomalus* Matα2 has Tup1 and Mcm1-interaction regions, it does not repress the **a**-specific genes and its DNA-binding domain is not capable of a-specific gene repression with Mcm1. This is most likely a derived trait found only in the lineage leading to *W. anomalus*.

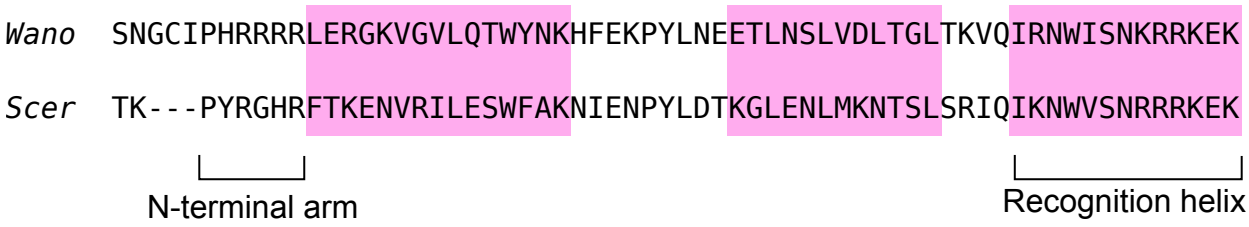


Figure 2. Alignment of *S. cerevisiae* and *W. anomalus* Matα2 homeodomains.

Alignment of *S. cerevisiae* ("*Scer*") and *W. anomalus* ("*Wano*") Matα2 DNA-binding domains constructed with MUSCLE. The two sequences are identical in 25/60 (40%) residues. The three alpha helices in the homeodomain three helix bundle are marked with pink boxes. The two regions that directly contact DNA (the N-terminal arm and the recognition helix) are highlighted. An alignment of the Matα2 DNA-binding domain from additional Saccharomycotina species is shown in Chapter 2.

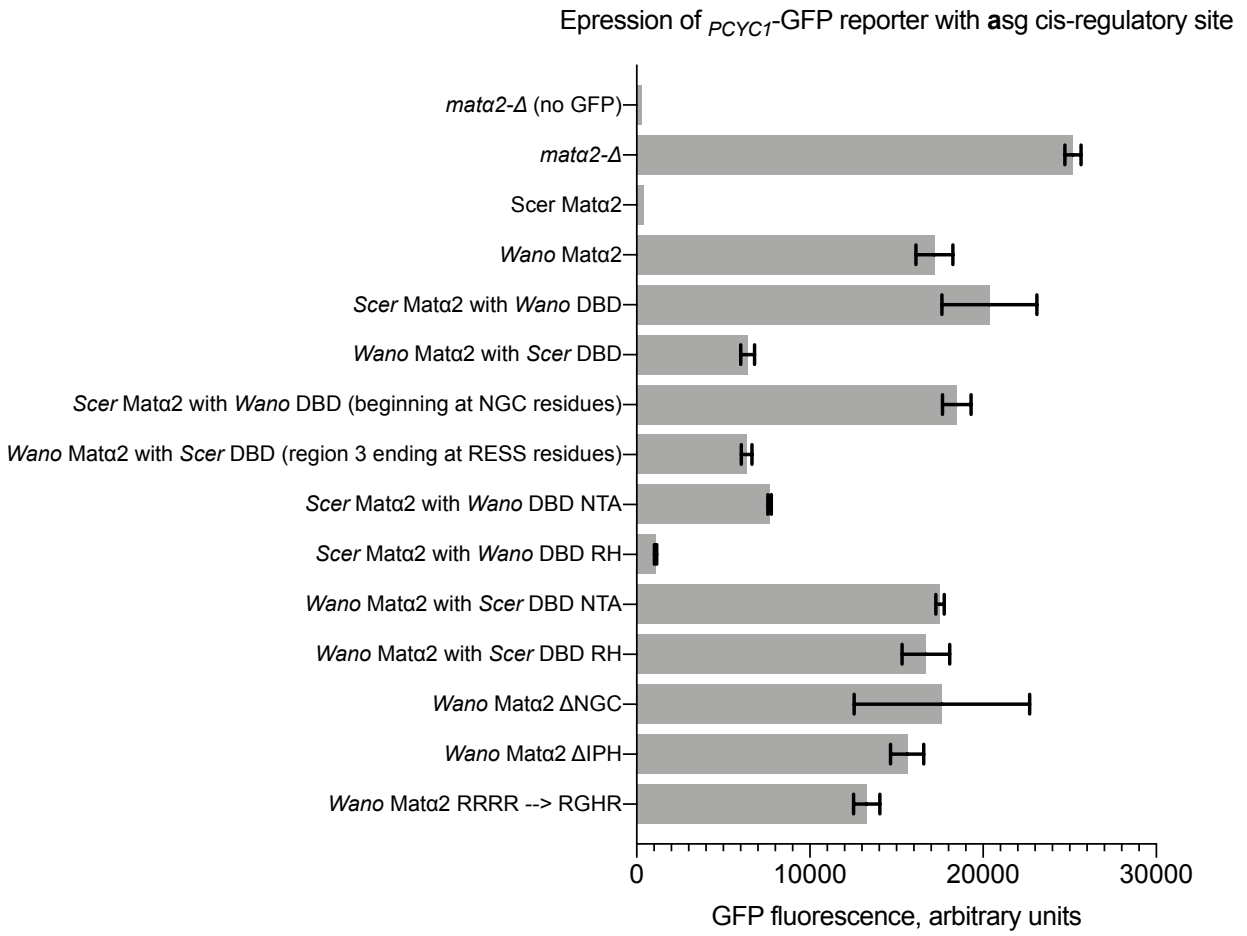


Figure 3. N-terminal arm is greater contributor to *W. anomalus* Matα2 DNA-binding domain's inability to repress a**-specific gene transcription**

The **a**-specific gene *cis*-regulatory sequence from the *S. cerevisiae STE2* gene was placed into a P_{CYC1}-GFP reporter construct and used to measure the ability of *S. cerevisiae* and *W. anomalus* Matα2s to repress **a**-specific gene transcription. *MATα2* constructs were inserted into a *S. cerevisiae* α cell with *MATα2* deleted (*matα2-Δ*) and driven by the *S. cerevisiae MATα2* promoter. In addition, portions of the *W. anomalus* Matα2 DNA-binding domain were swapped into the *S. cerevisiae* Matα2 protein and tested for their ability to repress **a**-specific gene transcription. "DBD: indicates the DNA-binding domain, "RH" indicates the recognition helix of DNA-binding domain, "NTA" indicates the N-terminal arm of the DNA-binding domain, and other abbreviations refer to specific amino acids within the Matα2 sequence. Mean and SD of three independent genetic isolates grown and tested in parallel are shown.

Strain number (yCSB #)	Genotype
120 A	*MATα matα2Δ::KanMX*
197	*MATα matα2Δ::KanMX ura3Δ::hph-$P_{CYC1\ (Scer\ STE2\ CRE)}$-GFP*
200	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-Scer MATα2 ura3Δ::hph-$P_{CYC1\ (Scer\ STE2\ CRE)}$-GFP*
312	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-Wano MATα2 ura3Δ::hph-$P_{CYC1\ (Scer\ STE2\ CRE)}$-GFP*
273	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Scer sequence with Wano region 4) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
274	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with Scer region 4) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
304	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Scer sequence with Wano region 4 beginning at NGC residues) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
305	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with Scer region 4, region 3 ending at RESS residues) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
296	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Scer sequence with Wano DBD NTA) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
297	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Scer sequence with Wano DBD recognition helix) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
323	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with Scer DBD NTA) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
324	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with Scer DBD recognition helix) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
315	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with ΔNGC residues in DBD NTA) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
316	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with ΔIPH residues in DBD NTA) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*
317	*MATα matα2Δ::KanMX TRP1::$P_{MATα2}$-MATα2 (Wano sequence with DBD NTA sequence changed from RRRR to RGHR) ura3Δ::hph-$P_{CYC1\ Scer\ STE2\ CRE)}$-GFP*

Table 1. *S. cerevisiae* strains used in this chapter
All strains are in the W-303 background, genotypes below denote modifications to the W-303 background (i.e., not recounting all auxotrophies that remained in the background). Note: "NTA" refers to the N-terminal arm of the Matα2 DNA-binding domain, and "DBD" means DNA-binding domain. "NGC," "IPH," "RRRR," and "RGHR" represent those amino acid sequences in the Matα2 sequence.

Materials and Methods

Growth conditions and media

All strains were grown on yeast extract peptone dextrose media (YEPD) at 30°C.

Strain construction in *S. cerevisiae*

All *S. cerevisiae* experiments were performed in the W303 strain, and transformations were performed by a published lithium acetate/single-stranded carrier DNA/polyethylene glycol method (Gietz and Woods, 2002). *MATα2* was deleted from the W303 α cell using a KanMX marker amplified with homology to *MATα2*.

MATα2 expression constructs for **a**-specific gene transcriptional repression assays were cloned into pNH 604, a plasmid that integrates into the *S. cerevisiae* genome in single copy at the *trp1* locus using a *TRP1* marker (Baker et al., 2012; Youk and Lim, 2014). pNH 604 was digested with SacII and KpnI to excise the Tef overexpression promoter. *MATα2* constructs were amplified from cDNA or synthesized as gBlock synthetic gene fragments (Integrated DNA Technologies) and fused to the *S. cerevisiae MATα2* promoter by fusion PCR, then inserted into pNH 604 between the SacII/KpnI restriction sites. pNH 604 were linearized by PmeI digestion, transformed into the *MATα matα2-Δ* strain, and selected for on SD-trp plates.

The P_{CYC1}-GFP reporter was designed to integrate in single copy at the *ura3* locus using the *hph* hygromycin resistance marker. It was adapted from the GFP reporter pTS 61 and the beta-galactosidase reporter pLG699z (Guarente and Ptashne, 1981; Sorrells et al., 2015). P_{CYC1} was inserted upstream of GFP, with the addition of a KpnI restriction site, allowing cis-regulatory elements to be inserted between the XhoI and KpnI sites in P_{CYC1}, between the upstream activating sequence and transcription start site of the promoter. The hph marker is up stream of P_{CYC1}-GFP, and the two together are flanked by homology to *URA3*. Overlapping oligos with the **a**-specific

gene cis regulatory elements and phosphorylated sticky ends were designed, annealed, and ligated into the XhoI/KpnI sites.

Design of *MATα2* mutant alleles

To construct chimeric alleles of *MATα2*, we defined the homeodomain the Matα2 protein based on the *S. cerevisiae* sequence as previously described (Baker et al., 2012). The DNA-binding domain is region 4, or amino acids numbered 128-188 in the *S. cerevisiae* sequence.

GFP reporter assays

Single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in SD supplemented with amino acids and uracil, and then grown for 4-4.5 hours at 30°C. GFP fluorescence was measured on a BD LSR II flow cytometer. 10,000 cells were measured per sample in each experiment. Cells were gated to exclude debris, and the mean GFP fluorescence was calculated for each reporter strain sample. Three independent genetic isolates of each reporter strain were measured.

References

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

Derelle, R., Lopez, P., Le Guyader, H., and Manuel, M. (2007). Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. Evol. Dev. *9*, 212–219.

Gietz, R.D., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Meth. Enzymol. *350*, 87–96.

Guarente, L., and Ptashne, M. (1981). Fusion of Escherichia coli lacZ to the cytochrome c gene of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.a. *78*, 2199–2203.

Li, T., Jin, Y., Vershon, A.K., and Wolberger, C. (1998). Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. Nucleic Acids Res. *26*, 5707–5718.

Li, T., Stark, M.R., Johnson, A.D., and Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. Science *270*, 262–269.

Mathias, J.R., Zhong, H.L., Jin, H.H., and Vershon, A.K. (2001). Altering the DNA-binding specificity of the yeast Mat alpha 2 homeodomain protein. J. Biol. Chem. *276*, 32696–32703.

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

Starr, T.N., and Thornton, J.W. (2016). Epistasis in protein evolution. Protein Science *25*, 1204–1218.

Tan, S., and Richmond, T.J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature *391*, 660–666.

Tuch, B.B., Galgoczy, D.J., Hernday, A.D., Li, H., and Johnson, A.D. (2008a). The evolution of combinatorial gene regulation in fungi. PLoS Biol *6*, e38.

Tuch, B.B., Li, H., and Johnson, A.D. (2008b). Evolution of eukaryotic transcription circuits. Science *319*, 1797–1799.

Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., and Pabo, C.O. (1991). Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell *67*, 517–528.

Youk, H., and Lim, W.A. (2014). Secreting and Sensing the Same Molecule Allows Cells to Achieve Versatile Social Behaviors. Science *343*, 1242782–1242782.

(2011). A Handbook of Transcription Factors (Dordrecht: Springer Netherlands).

Chapter Four

Wickerhamomyces anomalus cell-type specific genes and their regulation

Introduction

Cell-type specific gene regulation in the Saccharomycotina yeast has proven to be a fruitful model for studying the ways in which transcriptional circuits rewire (Dalal and Johnson, 2017; Nocedal and Johnson, 2015; Sorrells and Johnson, 2015). The **a**, α, and **a**/α cell types are ancient features of Saccharomycotina biology that are defined by the expression of small sets of genes (typically numbering between five and twenty) in the different mating cell types. That is, **a**-specific genes are expressed only in **a** cells, α-specific genes are expressed only in α cells, and the haploid-specific genes are expressed in both **a** and α cells, but not in **a**/α cells. Members of the Johnson Lab have made great progress in working to understand details of their regulation in a mechanistic way and make inferences about the evolution of the mating-type transcriptional regulatory circuits, transcriptional regulatory proteins, and *cis*-regulatory sequences.

In the course of these investigations of regulatory circuits, an additional evolutionary pattern has emerged: the identities of some of the cell-type specific genes themselves also change over time. The **a**, α and haploid-specific genes each have a core set of deeply conserved genes that tend to be widely conserved in the Saccharomycotina yeasts, and are therefore likely were **a**, α, or haploid-specific in the last common ancestor of the Saccharomycotina. For example, the **a**-specific genes almost always include *STE2* and *MFA*, the receptor for the α-mating pheromone and the gene for the **a**-mating pheromone, respectively (Baker et al., 2012). In turn, the α-specific genes always include *STE3* and *MFα*, the receptor for the **a** mating pheromone and the α-mating pheromone gene, respectively (Baker et al., 2011). The haploid-specific genes always include *STE4*, *STE18*, and *GPA1*, genes encoding the heterotrimeric G protein necessary for initiating mating (Booth et al., 2010). The functions of these deeply conserved "core" cell-type specific genes are essential for producing cell-type specific mating behaviors. Because their expression is

the defining feature of the different cell types, there is a logic to the ancient conservation of their expression patterns.

Nonetheless, not all cell-type specific genes are deeply conserved in this way. In particular, members of the **a** and haploid-specific genes can be specific to individual species and lineages (Baker et al., 2012; Booth et al., 2010). While any given Saccharomycotina yeast with **a** and α mating cell types likely shares the core of deeply conserved cell-type specific gene, there may also be genes brought into the networks, or lost from the network in that yeast's lineage alone. As discussed previously (Chapter 1), even simple gene regulatory changes such as the gain or loss of a single *cis*-regulatory sequence that changes the regulation of one gene product can have a substantial effect on an organism's morphological and physiological traits, and thus matter for its evolution. Whether this is also the case for some of the lineage-specific cell-type specific genes in yeast has never been tested. For example, it could be that some of the **a**-specific genes that are only **a**-specific in certain lineages also have **a**-specific mating phenotypes in those lineages.

In my previous studies examining the regulation of the **a** and haploid specific genes in *Wickerhamomyces anomalus*, I primarily focused on the ancient cell-type specific genes as models to exemplify the overall gene expression patterns (Chapter 2). I did this to simplify the analysis, to rely on genes with annotations and homology we can have a high level of confidence in, and because the actual identities of the genes are not important for understanding the circuitry of their regulation. Here, I will instead investigate the cell-type specific genes in *W. anomalus* in an unbiased manner. In particular, what are all of the **a**, α, and haploid-specific genes in *W. anomalus*? Are there any cell-type specific genes in *W. anomalus* that have not been described in other lineages? And finally, what can we surmise about their regulation, and can we detect any new instances of transcriptional circuit rewiring?

Results

Identification of the Phaffomycetaceae mating-type loci and mating-type transcriptional regulators

As described in previously, the genome sequenced *W. anomalus* strain (NRRL Y-366-8) is an **a** cell (see Chapter 2 Methods). The **a**-mating type locus (*MAT**a***) encodes the **a**-mating type transcriptional regulators *MAT**a**1* and *MAT**a**2*, and is located between the *DIC1* and *SLA2* genes, which are syntenic with the mating type locus in many Saccharomycotina species (Riley et al., 2016; Wolfe and Butler, 2017). Unlike species like *S. cerevisiae* and *K. lactis* (the Saccharomycetaceae clade), *W. anomalus* and other yeast in the Phaffomycetaceae are not known to undergo mating-type switching and there is no evidence for a silenced mating-type locus in their genomes (Kurtzman and Fell, 1998; Wolfe and Butler, 2017). Therefore, a *W. anomalus* **a** cell alone is insufficient to determine the sequence of the mating type transcriptional regulators from the α-mating type locus. To complete our understanding of the *W. anomalus* mating-type loci, Trevor Sorrells previously determined the sequence of the α-mating type locus (*MATα*) by designing oligonucleotides to PCR amplify the region between the *DIC1* and *SLA2* coding sequences of NRRL Y-2153-4, another *W. anomalus* isolate that has been reported to mate with the sequenced **a** cell. This sequencing revealed the genes for the α-mating type transcriptional regulators *MATα1* and *MATα2*. All of the *W. anomalus* mating transcriptional regulator genes except for *MATα1* contain introns, so we amplified and sequenced *MATα2*, *MAT**a**1* and *MAT**a**2* from cDNA to determine their exon coding sequences.

Identification of Mata**1 and Matα2-regulated genes in *W. anomalus* a/α cells**

As previously described, we tested whether Matα2 and Mat**a**1 are required for regulation of the haploid-specific genes in *W. anomalus* by deleting *MAT**a1*** and *MATα2* from **a**/α cells and measuring changes in gene expression patterns genome-wide by mRNA-seq (Chapter 2). Upon deletion of either *MAT**a1*** or *MATα2*, the haploid-specific genes are de-repressed, showing that Mat**a**1 and Matα2 are indeed required for haploid-specific gene repression in **a**/α cells (Figure 1). In addition, the **a** and α-specific genes also increase in expression, due to the de-repression of their direct activators, Mat**a**2 and Matα1.

To determine the broader set of Mat**a**1-Matα2 regulated genes in *W. anomalus*, I used the differential expression analysis software sleuth to identify the transcripts that are greater than 3-fold differentially expressed at a false discovery rate $q < 0.05$ in both deletions strains (*MAT**a**/MATα mata1*-Δ and *MAT**a**/MATα matα2*-Δ) with respect to a wildtype **a**/α cell (*MAT**a**/MATα*) (Bray et al., 2016; Pimentel et al., 2017). 47 transcripts met these criteria, including transcripts for the expected haploid, **a**, and α-specific genes (Table 1). *MAT**a2*** and *MATα1*, as well as *Nat* (the selectable marker used to delete *MAT**a1*** and *MATα2*) also met these criteria and were excluded from further analysis.

To ask whether these Mat**a**1 and Matα2-regulated genes are haploid, **a**, or α-specific, I examined their relative expression levels from a separate mRNA-seq experiment in which I measured gene expression in wildtype *W. anomalus* **a**, α and **a**/α cells (Figure 2). In this experiment the **a** cell strain is the genome sequenced strain NRRL Y-366-8, the α cell strain is NRRL Y-2153-4, and the **a**/α cell strain is the mating product of the two (the same parent strain for the *MAT**a1*** and *MATα2* deletions described above).While the **a** and α strains are both assigned to the *W. anomalus* species and can successfully mate, they are not isogenic and have both genotypic and phenotypic differences. Strikingly, all of the haploid-specific genes are more highly expressed in

the α cell type than the **a** cell. This appears to be a constitutive feature of this α cell isolate and may perhaps be attributed to its elevated *STE12* expression level. In *S. cerevisiae,* Ste12 activates gene expression upon pheromone induction, and it may be this α cell (which is also constitutively filamentous) is constitutively in a pheromone-induction like state.

I categorized the 16 genes that are at least two-fold higher in both the **a** and α strains relative to the **a**/α strain as haploid-specific. The eight genes that are at least two-fold higher in the α strain than the **a**/α strain, but not two-fold higher in the **a** strain than in the **a**/α strain are α-specific (with the exception of *STE12*, as noted above). The 11 genes that are two-fold higher in the **a** strain than the **a**/α strain, but not two-fold higher in the α strain than the **a**/α strain are **a**-specific. Eight additional genes did not meet these criteria and are as yet uncategorized. These **a**, α, haploid, and uncategorized genes are all more highly expressed in **a**/α cells when *MAT**a**1* or *MATα2* are deleted (Figure 3). The uncategorized genes tend not to be the most highly de-repressed transcripts in the deletion strains, though they are consistently de-repressed at a level similar to most of the **a**-specific genes.

With a more lenient set of cut offs, the assignments of certain cell-type specific genes changes. For example, if we consider **a**-specific genes to be those that are at least 1.5-fold higher in the **a** strain than the **a**/α strain, but not 1.5-fold higher in the α strain than in the **a**/α strain (rather than the 2-fold cut offs described above), then *RAM2*, *MUM1*, and *MAD1* can be considered additional **a**-specific genes, rather than "uncategorized". For this reason and additional reasons explored in the discussion, these gene expression assignments should be treated as preliminary for the time being. Additional, independent ways of testing these gene expression patterns will be needed to make more definitive assignments. In addition, at least two genes that are likely truly cell-type specific are missing from these initial lists. First, the genes for the **a**-mating pheromone (*MFA*) are excluded from the mRNA-seq analysis because the transcript is too small to quantified

by the alignment software used here. In addition, the **a**-specific gene *ASG7* is also missing because it is so lowly expressed in non-**a** cells that it is discarded by the differential expression software. Nonetheless, I previously confirmed that both *MFA* and *ASG7* are indeed **a**-specific genes that are activated by Mat**a**2 using the NanoString nCounter system to measure transcript abundance instead of RNA-seq (Chapter 2). *MFA* and *ASG7* should be considered to be **a**-specific genes moving forward.

***W. anomalus* α-specific genes**

The *W. anomalus* α-specific genes as defined above are the receptor for the **a**-mating pheromone (*STE3*), two copies of the **a**-pheromone blocker *AFB1*, three copies of the α-mating pheromone (*MFα*), a peptidase required for α pheromone maturation (*STE13*), and the α-agglutinin *SAG1*. This set of genes is exactly conserved with *S. cerevisiae*, except for the fact that there are three copies of the α-mating pheromone gene and only two in *S. cerevisiae*, and two copies of *AFB1* and only one in *S. cerevisiae* (Galgoczy et al., 2004).

To determine which α-specific genes are directly activated by Matα1 I generated an α-specific gene *cis*-regulatory motif for the *W. anomalus* α-specific genes *de novo* and checked for its presence in the α-specific genes individually (Figure 4a). Strong Matα1-Mcm1 motifs are found in the upstream regions of all of the α-specific genes except for *STE13*, suggesting that all but *STE13* are directly activated by Matα1-Mcm1 (Figure 4b). This is likely similar to the case in *S. cerevisiae*, where *STE13* is not a Matα1-activated or α-specific gene, but is upregulated in expression by **a** pheromone (Achstetter, 1989). Its relative upregulation in the α cells in this analysis may be related to the constitutive pheromone-like activation seen in the haploid-specific genes and *STE12*.

In general, the identities of the *W. anomalus* α-specific genes are greatly conserved with that of *S. cerevisiae*. In addition, the *W. anomalus* α-specific gene *cis*-regulatory motif is nearly

identical with the *S. cerevisiae* Matα1-Mcm1 binding site. The DNA-binding specificity of Matα1 has co-evolved with its *cis*-regulatory site in some clades of the Saccharomycotina yeast (Baker et al., 2011). The great deal of similarity between the *W. anomalus* and *S. cerevisiae* α-specific gene *cis*-regulator sequences is evidence that these changes occurred in the lineage leading to *Candida albicans* rather than *S. cerevisiae*.

W. anomalus a**-specific genes**

The **a**-specific genes as defined above are *STE2*, *STE6*, *BAR1*, *STE14*, *RTA1*, *AMS1*, *DAN4*, transcript #61221 (gm1.1447_g), transcript #65667 (gm1.5893_g), transcript #98304 (fgenesh1_kg.8_#_102_#_isotig12366), and transcript #25600 (e_gw1.1.2203.1). Whereas the *W. anomalus* α-specific genes are exactly conserved with *S. cerevisiae*, the **a**-specific genes contain both conserved and non-conserved genes. Conserved **a**-specific genes include the gene for the receptor for the α-mating pheromone (*STE2*), a transporter required for export of the **a**-mating pheromone (*STE6*), and a protease that cleaves the α-mating pheromone, relieving **a** cells from pheromone-induced cell cycle arrest (*BAR1*). In *S. cerevisiae STE14*, like *STE13* in α cells, has an **a**-specific function in the **a** cell, but its expression is not actually **a**-specific. The rest of the candidate **a**-specific genes are not **a**-specific genes in other species.

As described in Chapter 2, I generated an **a**-specific gene *cis*-regulatory motif using sequences from *W. anomalus* and its sister species *W. ciferrii* (Figure 5a). I used this motif to score these candidate **a**-specific genes, and found that only *STE2*, *STE6*, *BAR1,* and *DAN4* had strong matches to the Mat**a**2-Mcm1 binding site (Figure 5b). Surprisingly given its expression pattern and strong match to an Mat**a**2-Mcm1 binding site, I previously found that *DAN4* expression in **a** cells does not depend on Mat**a**2 activation (data not shown). That experiment was performed in rich media (YEPD), so there may be an as yet undetected Mat**a**2-dependent effect on *DAN4* transcription in other conditions. It is not clear whether Mat**a**2 directly regulates the other

candidate **a**-specific genes that do not have strong matches to the Mat**a**2-Mcm1 binding sites and were not included in the NanoString nCounter experiment presented in Chapter 2.

***W. anomalus* haploid-specific genes**

The *W. anomalus* haploid-specific genes as defined above are *STE4*, *RME1*, *FAR1*, *FUS3*, *GPA1*, *STE18*, *AXL1*, *HAC1*, *FEN2*, *PIN2*, transcript #60982 (gm1.1208_g), transcript #65336 (gm1.5562_g), transcript #78766 (fgenesh1_pg.3_#_454), transcript #95749 (fgenesh1_kg.5_#_706_#_contig04848), transcript #84228 (fgenesh1_pm.3_#_426), and transcript #90217 (fgenesh1_kg.2_#_660_#_isotig10953). Of these, *STE4*, *RME1*, *FAR1*, *FUS3*, *GPA1*, *STE18*, and *AXL1* are all haploid-specific genes in *S. cerevisiae*, and *STE4*, *FAR1*, *GPA1*, and *STE18* are also haploid-specific in *C. albicans*. *STE4*, *STE18*, and *GPA1* encode the three components of the heterotrimeric G-protein involved in initiating the mating response. Together with *FAR1*, they are the so-called "core" haploid specific genes that are deeply conserved essential to the mating process (Booth et al., 2010).

Which of these potential haploid-specific genes are directly regulated by Mat**a**1-Matα2? I did not succeed in generating a *de novo cis*-regulatory motif from the haploid-specific gene promoter regions in *W. anomalus*, despite varying many parameters in the MEME motif generation software. Unable to derive an endogenous haploid-specific gene motif, I instead asked which candidate haploid specific genes are likely directly regulated by Mat**a**1 and Matα2 by using a previously described *S. cerevisiae* Mat**a**1-Matα2 binding motif to score the promoters of the candidate haploid-specific genes and all of the other Mat**a**1 and Matα2-regulated genes for the presence of possible Mat**a**1-Matα2 binding sites (Booth et al., 2010). Strikingly, few of these genes have sequences that match the Mat**a**1-Matα2 binding site from *S. cerevisiae*. The *RME1* promoter has two such Mat**a**1-Matα2 *cis*-regulatory sequences (see Chapter 2), and the *FAR1* promoter has one very strong Mat**a**1-Matα2 cis-regulatory sequence. In addition, *GPA1*, two of the *MFα* genes,

the TFIID subunit encoded by transcript #80487 (fgenesh1_pg.5_#_318), and the transcription factor encoded by transcript #65667 (gm1.5893_g) have relatively strong Mat**a**1-Matα2 sites (strong sites here defined as a score >10, where a maximum score is approximately 12, see file "2019-02-19 a1alpha2 pwm at a1alpha2genes.pdf" for complete analysis). Therefore, while over 40 genes are differentially regulated by Mat**a**1 and Matα2, and 16 of those meet the criteria to be haploid-specific as described here, it is not clear from bioinformatics alone which, if any, of the haploid-specific genes are directly bound by Mat**a**1-Matα2.

Not all haploid-specific genes are directly bound by Mata**1 and Matα2 in chromatin immunoprecipitation**

To ask which cell-type specific genes are directly regulated by Mat**a**1 and Matα2, I introduced N-terminal HA-tags into both transcription factors, inserted them into the *W. anomalus* genome, and performed chromatin immunoprecipitation. As described in Chapter 2, HA-Mat**a**1 and HA-Matα2 both reproducibly pull down the *RME1* promoter region, whether they are inserted into the **a** or α haploid or **a**/α diploid background. To ask what regions these proteins bind genome-wide, I also performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). I sequenced immunoprecipitated DNA from wildtype **a**/α cells without a HA-tagged gene, **a** cells with *HA-MATα2* added, and α cells with *HA-MAT**a**1* added. In the design of this experiment I did not control for the expression levels of the Mat**a**1-Matα2 target genes. That is, while the haploid-specific genes are repressed in the **a**/α cell, they are active in the **a** and α cells despite the addition of the tagged genes, because these tagged genes did not complement the wildtype alleles (data not shown). While I believe the results of the experiment are nonetheless interpretable, this is an important caveat to keep in mind. Surprisingly, I found very few highly enriched peaks were reproducibly called between all of the tagged samples compared to the untagged control. These peaks always included the *RME1* promoter region, in all pairwise

comparisons of tagged vs. untagged samples. In addition to *RME1*, approximately 6 additional peaks were frequently called (Table 2). These peaks contain no other haploid-specific genes, or any apparent Mat**a**1 and Matα2-regulated genes.

***AXL1*: a Mat**a**2-activated haploid-specific gene?**

In the present analysis, *AXL1* is categorized as a haploid-specific gene. However, in Chapter 2 I presented evidence that it is actually a Mat**a**2-activated **a**-specific gene that is more highly expressed in the **a** cell (yCSB 5) than the α cell I compared it to (yCSB 342) (see Chapter 2, Figure 2, Supplementary Figure 2). Indeed, it has a strong match to the Mat**a**2-Mcm1 *cis*-regulatory site (Figure 6). Surprisingly, in this analysis it is relatively highly expressed in the wildtype α cell (yCSB 4). I propose that *W. anomalus AXL1* is actually a Mat**a**2-activated gene that is also expressed in α cells. That is, it is a haploid-specific gene that in **a** cells is activated by Mat**a**2. As it does not have a strong Matα1-Mcm1 binding site, it is not clear what is regulating its expression in α cells.

Discussion

In this Chapter, I sought to analyze the effects of deleting *W. anomalus MAT**a**1* and *MATα2* from **a**/α cells in an agnostic way. My aim was to more completely describe the full set of cell-type specific genes in this species. While the α-specific genes are highly conserved with the α-specific genes in other Saccharomycotina yeast species, I found that the **a** and haploid-specific genes likely contain members that are thus far found only in the *W. anomalus* lineage. In addition, I show a lack of evidence for the direct regulation of most of the haploid-specific genes by Mat**a**1-Matα2, suggesting that there may be an additional intercalation of a transcription factor into the haploid-specific gene regulatory circuit in this lineage (Booth et al., 2010).

There are several limitations to the present analysis that can be addressed by additional experiments. First, the **a** and α cell strains that I used to sort genes in this analysis are not

genetically identical, and even show great differences in their basal expression levels of the haploid-specific genes. Creating a near-isogenic pair of **a** and α cells by repeatedly mating, sporulating, and backcrossing α progeny to the **a** type cell would provide an independent way to observe the normal expression patterns of these genes and greatly improve our ability to sort them. In addition, we could directly ask whether all of the candidate **a**-specific genes discussed in this chapter are indeed Mat**a**2-activated by measuring gene expression changes transcriptome-wide by mRNA-seq comparing a wildtype **a** strain to the **a** strain with *MAT**a**2* deleted. In addition, repeating the HA-Mat**a**1 and HA-Matα2 ChIP-seq experiments described here with additional controls would give us a more definitive sense of where in the genome Mat**a**1-Matα2 are directly binding. We could further test whether Mat**a**1-Matα2 repression of *RME1* is sufficient for repression of the haploid-specific genes in **a**/α cells by introducing a version of the *RME1* gene in which the promoter is not repressible by Mat**a**1-Matα2 (i.e., by deleting the necessary Mcm1-binding site described in Chapter 2) and measuring changes in gene expression. Finally, we could test the hypothesis that Rme1 and/or Ste12 are direct regulators of the haploids-specific genes instead of Mat**a**1-Matα2 by deleting them and measuring changes in haploid specific gene expression and mating ability. This and future analyses of the cell-type specific genes in *W. anomalus* may provide new case studies of the importance of gene regulatory evolution in the Saccharomycotina yeast.

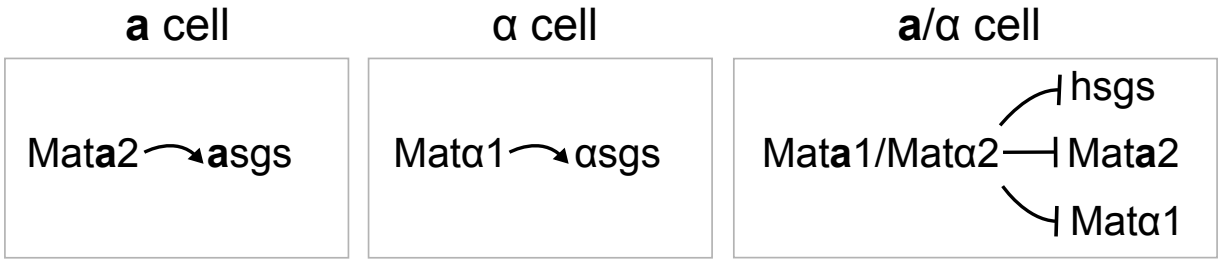


Figure 1. Cell-type specific gene regulation in *W. anomalus*

Summary of **a**, α, and haploid-specific gene regulation in *W. anomalus*. Mat**a**2 activates the **a**-specific genes in **a** cells, and Matα1 activates the α-specific genes in α cells. In **a**/α cells Mat**a**1and Matα2 repress the haploid specific genes, Mat**a**2 and Matα1, resulting in the repression of the **a** and α-specific genes in turn. This circuit summary is based on analysis of the deletions of *MAT**a**2*, *MAT**a**1*, and *MATα2* presented in Chapter 2.

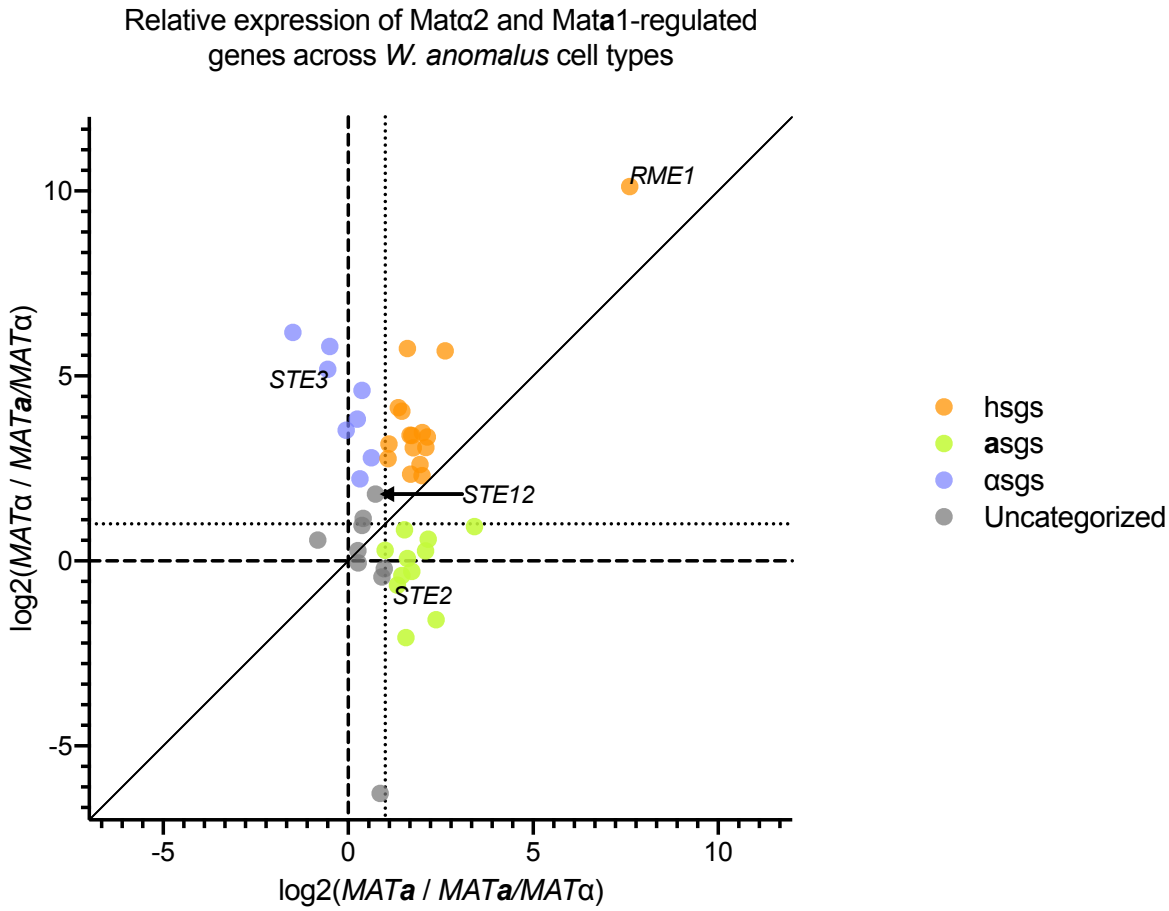


Figure 2. Relative expression of Matα2 and Mata1-regulated genes across wildtype *W. anomalus* cell types

Relative expression levels of genes that are differentially expressed upon deletion of either *MATα2* or *MAT**a**1* in an **a**/α cell in wildtype **a**, α, and **a**/α cells. Expression levels in transcripts per million were averaged from two independent samples of each cell type and used to calculate log2 ratios of expression levels between the haploid and **a**/α strains. The x axis shows the log2 ratio of expression in **a** cells compared to **a**/α cells, and the y axis shows the log2 ratio of expression in α cells compared to **a**/α cells. The smaller dashed lines indicate the threshold for a 2-fold effect size (i.e., 2-fold more highly expressed in the **a** or α strain relative to the **a**/α strain). Genes in the bottom right quadrant are at least 2-fold higher in **a** cells than in **a**/α cells but not in α cells, and are considered **a**-specific genes and are colored green. Genes in the upper left quadrant are at least 2-fold higher in α cells than in **a**/α cells, but not in **a** cells, are considered α-specific genes and are colored blue. Genes that do not meet the 2-fold threshold are uncategorized and colored gray. The haploid specific genes are at least 2-fold higher in both **a** and α cells relative to **a**/α cells and are colored orange. Notably, the haploid-specific genes are consistently higher in α cells than in **a** cells. This appears to be a constitutive effect in this natural α cell isolate and may be due to *STE12*'s elevated expression level. *STE12* is the only uncategorized gene that is 2-fold more highly expressed in the α cell but not in the **a** cell, but is not truly an α-specific gene. Labels for the **a**-specific gene *STE2*, the α-specific gene *STE3*, and the haploid-specific gene *RME1* are superimposed on their data points as representative cell-type specific genes.

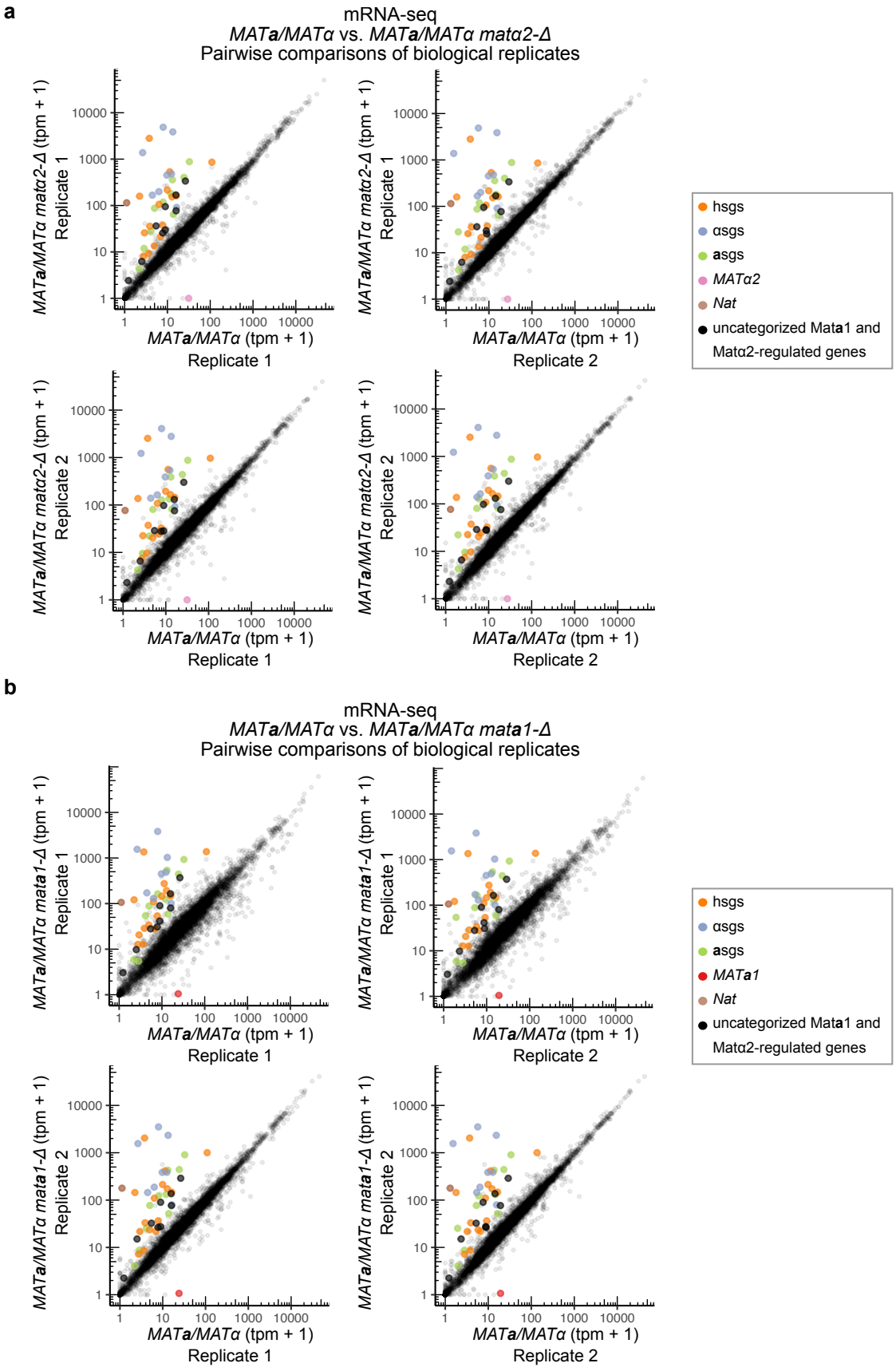
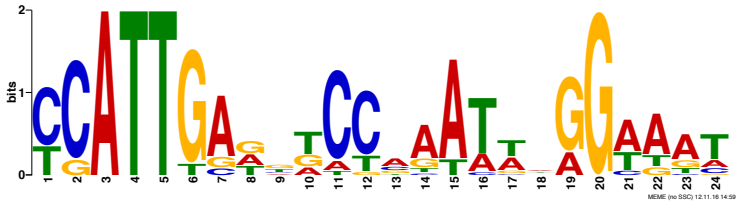

a
mRNA-seq
MATa/MATα vs. MATa/MATα matα2-Δ
Pairwise comparisons of biological replicates
MATa/MATα matα2-Δ (tpm + 1) Replicate 1
MATa/MATα (tpm + 1) Replicate 1
MATa/MATα matα2-Δ (tpm + 1) Replicate 1
MATa/MATα (tpm + 1) Replicate 2
MATa/MATα matα2-Δ (tpm + 1) Replicate 2
MATa/MATα (tpm + 1) Replicate 1
MATa/MATα matα2-Δ (tpm + 1) Replicate 2
MATa/MATα (tpm + 1) Replicate 2
1
10
100
1000
10000
hsgs
αsgs
asgs
MATα2
Nat
uncategorized Mata1 and Matα2-regulated genes
b
mRNA-seq
MATa/MATα vs. MATa/MATα mata1-Δ
Pairwise comparisons of biological replicates
MATa/MATα mata1-Δ (tpm + 1) Replicate 1
MATa/MATα (tpm + 1) Replicate 1
MATa/MATα mata1-Δ (tpm + 1) Replicate 1
MATa/MATα (tpm + 1) Replicate 2
MATa/MATα mata1-Δ (tpm + 1) Replicate 2
MATa/MATα (tpm + 1) Replicate 1
MATa/MATα mata1-Δ (tpm + 1) Replicate 2
MATa/MATα (tpm + 1) Replicate 2
hsgs
αsgs
asgs
MATa1
Nat
uncategorized Mata1 and Matα2-regulated genes

Figure 3. Expression of categorized Matα2 and Mata**1-regulated genes in** a/**α cells with *MATα2* or *MATa1* deleted**

a. mRNA-seq of wildtype *W. anomalus* **a**/α cells (*MAT**a**/MATα*) compared to **a**/α cells with *MATα2* deleted (*MAT**a**/MATα matα2-Δ*). Expression in shown in transcripts per million (tpm). *MATα2* expression is shown in pink, the **a**-specific genes as defined in this chapter are in green, the α-specific genes are in blue, the haploid-specific genes are in orange, the genes that are statistically significantly Mat**a**1-Matα2 regulated but not categorized are in opaque black, and the selective marker used to delete *MATα2* is in brown.

b. mRNA-seq of wildtype *W. anomalus* **a**/α cells (*MAT**a**/MATα*) compared to **a**/α cells with *MAT**a**1* deleted (*MAT**a**/MATα mat**a**1-Δ*). Mat**a**1-Matα2 regulated genes are labeled as described above, except *MAT**a**1* is labeled in red and *MATα2* is not labeled.

a



b

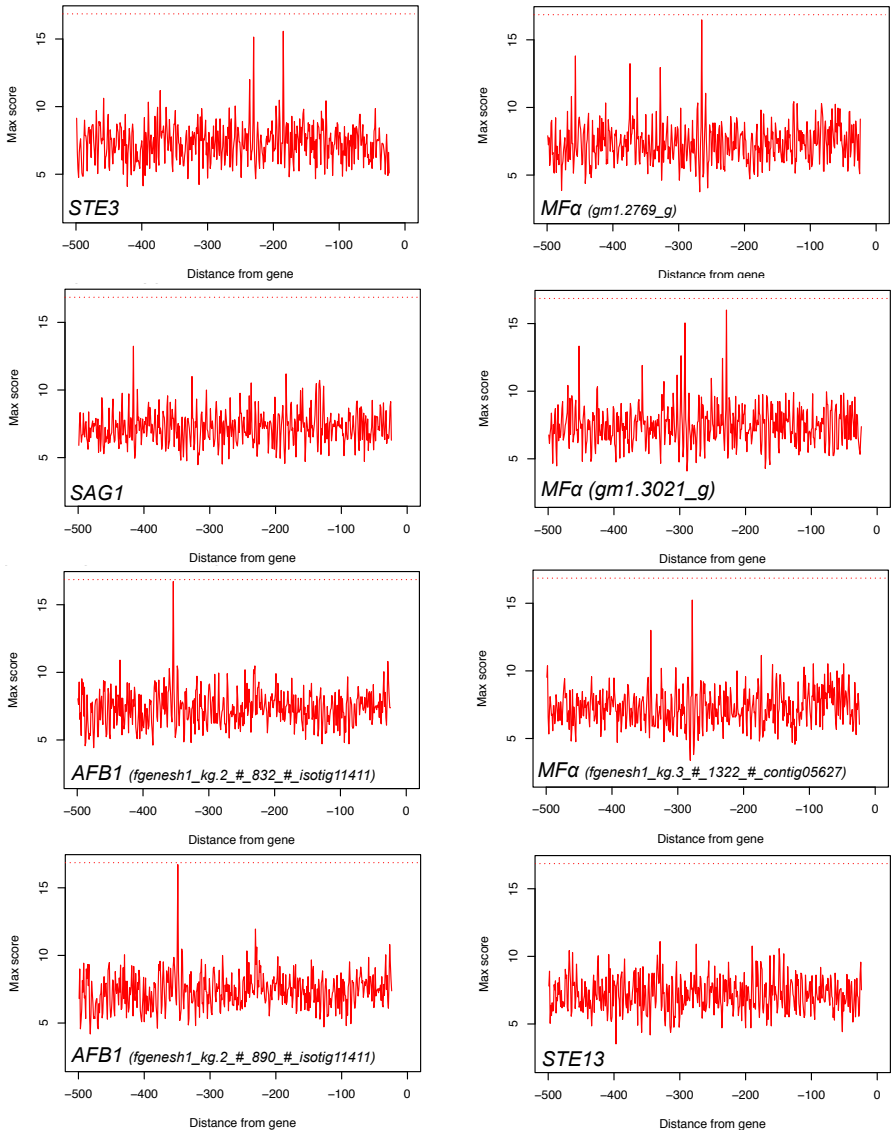
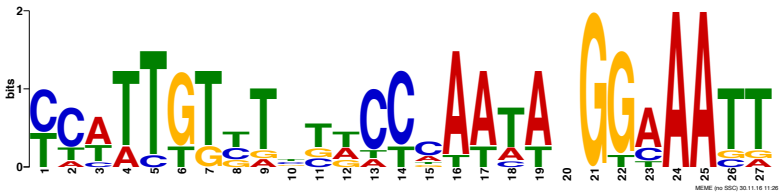


Figure 4. *W. anomalus* α-specific gene *cis*-regulatory sequences

a. *W. anomalus* Matα1-Mcm1 binding motif generated by MEME search for motifs between 20 and 30 nucleotides in length, occurring in any number of repeats in the 1000 basepairs upstream of *W. anomalus* α-specific gene coding sequences (Bailey et al., 2009).

b. 500 basepairs upstream of *W. anomalus* α-specific gene coding sequences scored for presence of the Matα1-Mcm1 motif in panel a. The dashed red line at the top of each plot indicates the maximum possible score for the Matα1-Mcm1 motif, and the solid red line indicates the score for how well every position in the region matches the Matα1-Mcm1 motif. All of the α-specific genes except for *STE13* have strong matches to the Matα1-Mcm1 binding sequence. Motif scores were calculated and plotted using the Biostrings R package, with code provided by Kyle Fowler.

a



b

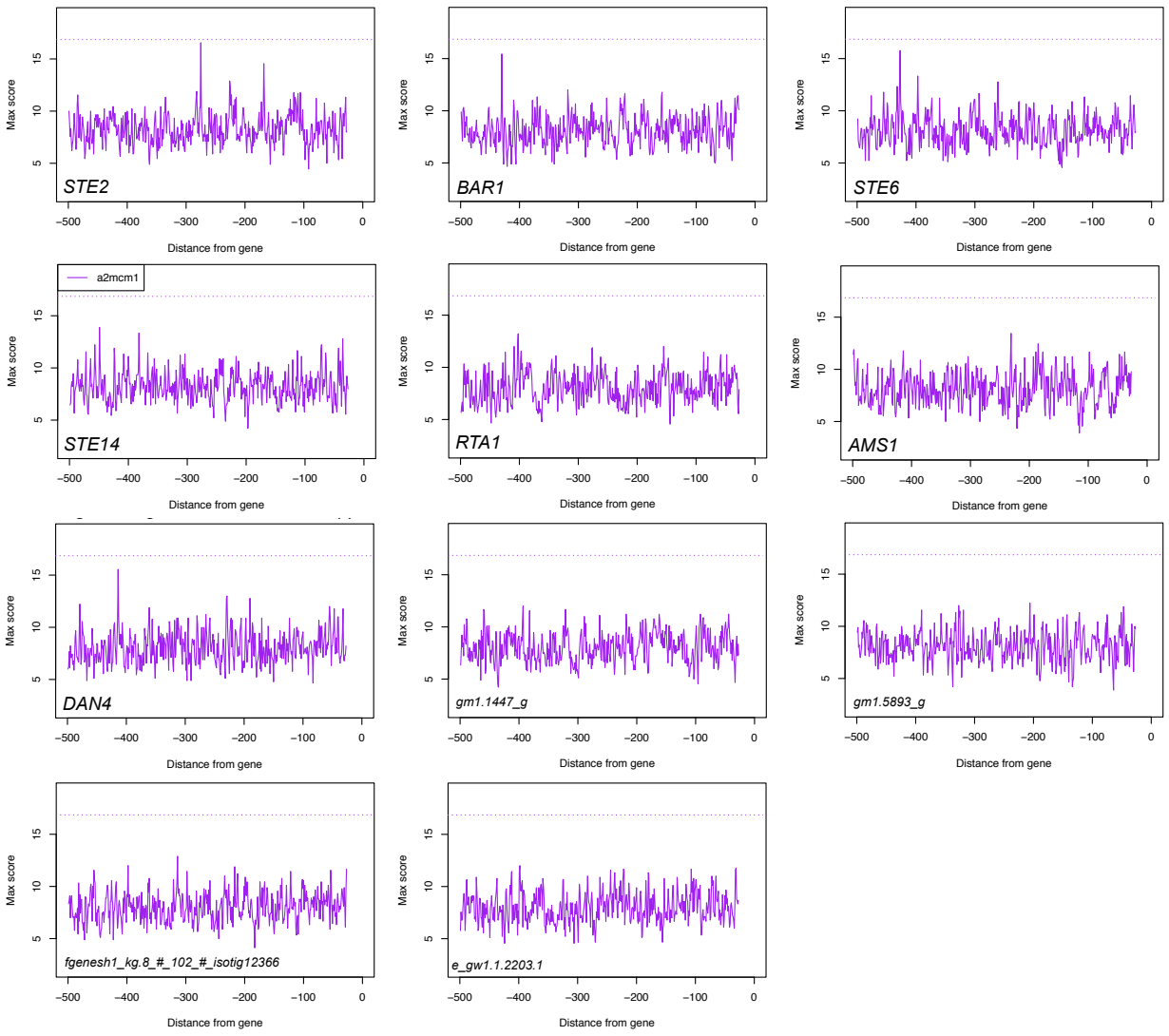


Figure 5. ***W. anomalus*** a**-specific gene** ***cis*****-regulatory sequences**

a. Mat**a**2-Mcm1 **a**-specific gene *cis*-regulatory motif generated by MEME search for motifs between 24 and 27 nucleotides in length occurring zero or one time in the sequences of candidate *W. anomalus* and *W. ciferrii* **a**-specific gene promoter regions (500 basepairs upstream of coding sequences).

b. *W. anomalus* **a**-specific gene promoter regions scored for presence of Mat**a**2-Mcm1 binding motif in panel a. Where gene annotations are uncertain, the JGI model name for each transcript is

indicated instead. The dashed purple line at the top of each plot indicates the maximum possible score for this motif, and the solid purple line indicates the score for how well every position in the region matches the Mat**a**2-Mcm1 motif. While *STE2*, *STE6*, and *BAR1* have strong matches to the Mat**a**2-Mcm1 binding sequence, the rest of these proposed **a**-specific genes do not. Motif scores were calculated and plotted using the Biostrings R package, with code provided by Kyle Fowler.

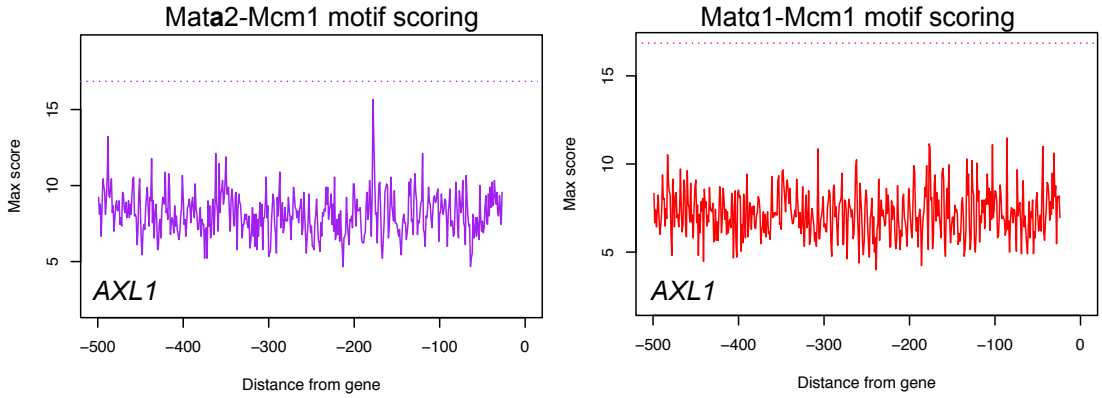


Figure 6. *AXL1* motif scoring

The *AXL1* promoter region has a strong Mata2-Mcm1 site, which may explain its Mat**a**2-dependent activation, despite its haploid-specific gene expression pattern.

Transcript (Transcript ID|Model Name)	Annotation	Cell-type specificity	Direct regulator
100850|estExt_fgenesh1_pm.C_4_t10078+B2:B46	*STE14*	**a**	
102672|estExt_fgenesh1_pg.C_1_t10079	*SAG1*	α	Matα1
17191|gw1.1.1201.1	*STE3*	α	Matα1
18095|gw1.1.1407.1	*STE2*	**a**	Mat**a**2
25109|e_gw1.1.93.1	*RAM1*	uncategorized	
25504|e_gw1.1.338.1	*STE13*	α	
25600|e_gw1.1.2203.1	an aldehyde dehydrogenase	**a**	
26608|e_gw1.1.2260.1	*STE4*	Haploid	
26943|e_gw1.1.1955.1	*GPA1*	Haploid	
28428|e_gw1.2.1586.1	*FUS3*	Haploid	
28696|e_gw1.2.1806.1	*STE12*	uncategorized	
29842|e_gw1.3.1722.1	*SST2*	uncategorized	
31720|e_gw1.4.1729.1	*AXL1*	Haploid	Mat**a**2
34379|e_gw1.5.215.1	*RTA1*	**a**	
50901|estExt_Genewise1Plus.C_1_t60080	*RAM2*	uncategorized	
53731|estExt_Genewise1Plus.C_3_t20389	*AMS1*	**a**	
59975|gm1.201_g	*RME1*	Haploid	Mat**a**1-Matα2
60982|gm1.1208_g	hypothetical protein	Haploid	
61221|gm1.1447_g	uORF	**a**	
61484|gm1.1710_g	*DAN4*	**a**	Mat**a**2, though not in YEPD?
62332|gm1.2558_g	*STE18*	Haploid	
62543|gm1.2769_g	*Mfα*	α	Matα1
62795|gm1.3021_g	*Mfα*	α	Matα1
65336|gm1.5562_g	uncharacterized protein	Haploid	
65667|gm1.5893_g	transcription factor	**a**	
65783|gm1.6009_g	*STE5*	uncategorized	
66009|gm1.6235_g	*MUM1*	uncategorized	
76275|fgenesh1_pg.1_#_375	*PIN2*	Haploid	
76783|fgenesh1_pg.1_#_883	*HAC1*	Haploid	
77692|fgenesh1_pg.2_#_342	*FAR1*	Haploid	
78766|fgenesh1_pg.3_#_454	Nucleolar GTPase/ATPase	Haploid	
80487|fgenesh1_pg.5_#_318	TFIID subunit	uncategorized	
80943|fgenesh1_pg.6_#_100	*BAR1*	**a**	Mat**a**2
82986|fgenesh1_pm.1_#_1010	*STE6*	**a**	Mat**a**2
83816|fgenesh1_pm.3_#_14	aquaporin	uncategorized	
84228|fgenesh1_pm.3_#_426	Choline transporter-like protein	Haploid	
84993|fgenesh1_pm.4_#_485	*FEN2*	Haploid	
85554|fgenesh1_pm.5_#_353	Fe2+/Zn2+ regulated transporter	uncategorized	
90217|fgenesh1_kg.2_#_660_#_isotig10953	a chitinase	Haploid	
90389|fgenesh1_kg.2_#_832_#_isotig11411	*AFB1*	α	Matα1
90447|fgenesh1_kg.2_#_890_#_isotig11411	*AFB1*	α	Matα1
92602|fgenesh1_kg.3_#_1322_#_contig05627	*Mfα*	α	Matα1
95749|fgenesh1_kg.5_#_706_#_contig04848		Haploid	
98304|fgenesh1_kg.8_#_102_#_isotig12366	a chitinase	**a**	
n/a	*MAT**a**2*		
n/a	*MATα1*		
n/a	*Nat*	n/a	n/a

Table 1. Complete list of Mata**1-Matα2 regulated genes in *W. anomalus* a/α cell**

The transcripts listed are greater than 3-fold differentially expressed in **a**/α cells with *MAT**a**1* or *MATα2* deleted, at a false discovery rate q < 0.05. Deletion of *MAT**a**1* or *MATα2* de-represses transcription of the cell-type specific activators *MAT**a**2* and *MATα1*, so this list also includes the **a** and α-specific genes. The cell-type specificity as assigned in Figure 2 is listed in the third column, and the presumed direct regulator of these genes (if inferable from *cis*-regulatory sites, and, in the case of *RME1*, ChIP-seq) is listed in the fourth column.

Scaffold #	Start point (bp)	End point (bp)	Nearest annotated gene model	Annotation
scaffold_8	395364	403275	fgenesh1_pg.8_#_184	tRNA deacylase
scaffold_7	455429	456126	fgenesh1_kg.7_#_400_#_isotig13249	?
scaffold_5	824577	824931	fgenesh1_pm.5_#_261	Aldehyde dehydrogenase
scaffold_8	421401	421812	gm1.6175_g	uncharacterized protein
scaffold_1	423903	424898	gm1.201_g	*RME1*
scaffold_6	217283	217599	fgenesh1_pg.6_#_83	Predicted hydrolase/acyltransferase
scaffold_2	637217	637465	estExt_Genemark1.C_2_t10297	Phospholipid-translocating P-type ATPase, flippase

Table 2. Regions reproducibly enriched in HA-Mata1 and HA-Matα2 chromatin immunoprecipitation

Strain number (yCSB #)	USDA ARS Strain number	Genotype
4	NRRL Y-2153-4	*MATα* Arg-
5	NRRL Y-366-8	*MAT**a***
8	n/a	*MAT**a**/MATα*
340	n/a	*MAT**a**/MATα matα2Δ-3::Nat*
351	n/a	*MAT**a**/MATα mat**a**1Δ::Nat*
410	n/a	*MAT**a** HA-MATα2::Nat*
411	n/a	*MATα HA-MAT**a**1::Nat*
412	n/a	*MAT**a**/MATα HA-MATα2::Nat*
413	n/a	*MAT**a**/MATα HA-MAT**a**1::Nat*

Table 3. *W. anomalus* strains used in this chapter

Materials and Methods

Growth conditions and media

All strains were grown on yeast extract peptone dextrose media (YEPD) at 30°C unless otherwise noted.

Strain construction in *W. anomalus*

The genome-sequenced strain of *W. anomalus*, NRRL Y-366-8, is also the *W. anomalus* type strain and an **a** cell. To generate an **a**/α strain of *W. anomalus*, NRRL Y-366-8 was plated on media containing 5-FOA to select for spontaneous uracil auxotrophs. An **a** cell auxotrophic for uracil was then crossed to NRRL Y-2153-4, an α cell with a naturally occurring arginine auxotrophy. The resulting **a**/α cell (yCSB 8) was selected for by its ability to grow on synthetic defined media with both arginine and uracil dropped out. The 2N diploid ploidy of yCSB 8 was confirmed by staining with SYTO™ 13 Green Fluorescent Nucleic Acid Stain followed by flow cytometry, and the presence of both *MAT**a*** and *MATα* mating type loci in yCSB 8 was confirmed by PCR. *MAT**a**/MATα* deletions were made in yCSB 8.

W. anomalus genes were deleted using a nourseothricin-resistance marker and a freeze-thaw transformation protocol optimized for use in *W. ciferrii* by Schorsch *et al*. (Schorsch et al., 2009). To construct deletion cassettes, we cloned homology arms into pCS.ΔLig4, a plasmid generated by Schorsch *et al*. (Schorsch et al., 2009). pCS.ΔLig4 contains *nat*, the nourseothricin-resistance gene, codon-optimized for expression in *W. ciferrii*, and under control of *W. ciferrii* P_{PDA1} and T_{TEF1}. It also contains a region of the *W. ciferrii LIG4* gene flanked by loxP sites, for inserting pCS.ΔLig4 into *LIG4* by ends-in recombination. To use the *W. ciferrii*-optimized *nat* marker to delete genes in *W. anomalus* we excised the *LIG4* fragment from pCS.ΔLig4 and ligated in 5' homology regions in between the SexAI and SalI restriction sites, and 3' homology regions between the SacI and SpeI restriction sites. Homology regions ranged from approximately 700

basepairs to 2000 basepairs in length. Transformed yCSB 8 cells were selected on 400 µg/ml nourseothricin. Deletions were confirmed by PCR testing for the absence of the deleted ORF, as well as the presence of the junctions between the deletion cassette and the surrounding genome.

Gene expression profiling by mRNA-seq

For gene expression profiling by mRNA-seq, single colonies were grown to saturation overnight in YEPD, diluted to OD_{600} = 0.1 in fresh YEPD, and then grown for 4-4.5 hours at 30°C. Cells were pelleted by centrifugation at 3000 rpm for 10 minutes, then flash frozen in liquid nitrogen and stored at -80°C. RNA was extracted, purified, and DNase-treated using an Ambion RiboPure RNA Purification kit for yeast. Quality of total RNA was evaluated on an Agilent Bioanalyzer using an Agilent RNA 6000 Pico Kit. For the gene deletion (*MAT**a**/MATα mata1*-Δ and *MAT**a**/MATα matα2*-Δ) compared to wildtype **a**/α cells (*MAT**a**/MATα*) experiment two rounds of polyA selection were performed using the Qiagen Oligotex mRNA Mini Kit, and mRNA quality was evaluated on an Agilent Bioanalyzer using an Agilent RNA 6000 Pico Kit. mRNA samples were then concentrated using a Zymo RNA Clean and Concentrator Kit. cDNA synthesis and library preparation were performed using a NEBNext Directional RNA-seq kit. Libraries were sequenced using single end 50 basepair reads on the Illumina HiSeq 4000 sequencing system in the UCSF Center for Advanced Technologies. For the wildtype *W. anomalus* **a**, α, vs. **a**/α mRNA-seq experiment, purified RNA was sent to the Columbia Genome Center where it was polyA selected, prepared into a sequencing library, and sequenced with 100 basepair paired-end reads (https://systemsbiology.columbia.edu/genome-center).

mRNA-seq analysis

Reads were pseudoaligned to all *W. anomalus* coding sequences using kallisto, and differential expression analyses were performed using sleuth(Bray et al., 2016; Pimentel et al., 2017). Coding sequences were downloaded from the Joint Genome Institute annotation of genes

for *Wickerhamomyces anomalus* NRRL Y-366-8 v1.0 (Riley et al., 2016). For differential expression analyses, we used the sleuth likelihood ratio test to compare the two deletion strains (*MAT**a**/MATα mata1*-Δ and *MAT**a**/MATα matα2*-Δ) a wildtype **a**/α cell (*MAT**a**/MATα*) and chose a false discovery rate cut off of $q < 0.05$ to identify differentially expressed genes. To further limit the stringency of our search, we focused on genes that were greater than three-fold differentially expressed in the two deletion strains with respect to the wildtype **a**/α strain.

Cis-regulatory motif discovery and scoring

A *cis*-regulatory motif for the Phaffomycetaceae **a**-specific genes was generated using MEME(Bailey et al., 2009). We used sequences 500 basepairs upstream of the **a**-specific gene orthologs in *W. anomalus* and *W. ciferrii* and used MEME to search for motifs between 24 and 27 bases wide occurring zero or one time in those sequences. These sequences were from the *W. ciferrii STE2*, *STE6*, and *AXL1* genes, and the *W. anomalus STE2*, *STE6*, *ASG7*, *AXL1*, *BAR1*, and *STE14* genes. This returned a motif that resembled a Mat**a**2-Mcm1 binding site. We also generated a *cis*-regulatory motif for the α-specific genes using a MEME search for motifs between 20 and 30 nucleotides in length, occurring in any number of repeats in the 1000 basepairs upstream of *W. anomalus* α-specific gene coding sequences. We scored the presence of *W. anomalus* Matα1, *W. anomalus* Mat**a**2, and *S. cerevisiae* Mat**a**1-Matα2 *cis*-regulatory motifs using the Biostrings R package, with code provided by Kyle Fowler.

Chromatin-immunoprecipitation

*HA-MAT**a**1* and *HA-MATα2* constructs were inserted into the *W. anomalus* genome via random integration. Chromatin immunoprecipitation (ChIP) was performed as previously described(Nocedal et al., 2017; Sorrells et al., 2015). Briefly, single colonies were grown to saturation overnight in YEPD, diluted to $OD_{600} = 0.1$ in fresh YEPD, then grown at 30°C to $OD_{600} = 0.4$. Cells were then crosslinked with formaldehyde, quenched with glycine, washed with TBS

(20mMTris/HCl pH 7.4, 150 mM NaCl), flash frozen in liquid nitrogen, and stored at -80°C. Cells were lysed at 4°C in lysis buffer (50 mM HEPES/KOH pH 7.5, 140 mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-Deoxycholate) plus protease inhibitors (Roche Complete Protease Inhibitor Cocktail EDTA-free) with 0.5 mm glass beads on a vortex for two hours. Recovered chromatin was then sheared via sonication on a Diagenode Biorupter (three ten minute periods of sonication on level 5 for 30 seconds, then off for 1 minute). 10 mg of anti-HA 12CA5 from Roche was used to bind the HA-tagged transcription factors overnight at 4°C in lysis buffer plus protease inhibitors, and protein-G sepharose beads were used to collect immunoprecipitated DNA. Immunoprecipitated DNA was then recovered by washing with lysis buffer, wash buffer (10mM Tris/HCl pH 8.0, 250mM LiCl, 0.5% NP-40, 0.5% Na-Deoxycholate, 1mM EDTA), and TE, then collected in elution buffer (50mM Tris/HCl pH8.0, 10mM EDTA, 1%SDS). Crosslinking was reversed by Proteinase K treatment for two hours at 37°C, then16 hours at 65°C. DNA was then cleaned up using the Qiagen MinElute kit. Library preparation was performed using a NEBNext Directional RNA-seq kit, and libraries were sequenced using single end 50 basepair reads on the Illumina HiSeq 4000 sequencing system in the UCSF Center for Advanced Technologies.

ChIP-seq data was analyzed as previously described, with a few modifications (Nocedal et al., 2017). Briefly, reads were aligned to the *W. anomalus* genome using bowtie, and peak characteristics were determined using SPP (Bailey et al., 2013; Feng et al., 2012). Peaks were called using MACS2 with the modification of keeping all duplicate reads. Duplicate reads were kept due to the saturation of the coverage of the genome in this sequencing experiment, and the relative scarcity and high enrichment of peaks.

References

Achstetter, T. (1989). Regulation of Alpha-Factor Production in Saccharomyces-Cerevisiae - a-Factor Pheromone-Induced Expression of the Mf-Alpha-1 and Ste13 Genes. Mol. Cell. Biol. *9*, 4507–4514.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput. Biol. *9*, e1003326.

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

Baker, C.R., Tuch, B.B., and Johnson, A.D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc. Natl. Acad. Sci. U.S.a. *108*, 7493–7498.

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

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Dalal, C.K., and Johnson, A.D. (2017). How transcription circuits explore alternative architectures while maintaining overall circuit output. Genes Dev. *31*, 1397–1405.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728–1740.

Galgoczy, D.J., Cassidy-Stone, A., Llinás, M., O'Rourke, S.M., Herskowitz, I., DeRisi, J.L., and Johnson, A.D. (2004). Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.a. *101*, 18069–18074.

Kurtzman, C., and Fell, J.W. (1998). The Yeasts - A Taxonomic Study (Elsevier).

Nocedal, I., and Johnson, A.D. (2015). How Transcription Networks Evolve and Produce Biological Novelty. Cold Spring Harbor Symposia on Quantitative Biology *80*, 265–274.

Nocedal, I., Mancera, E., and Johnson, A.D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. Elife *6*, 1.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690.

Riley, R., Haridas, S., Wolfe, K.H., Lopes, M.R., Hittinger, C.T., Goeker, M., Salamov, A.A., Wisecaver, J.H., Long, T.M., Calvey, C.H., et al. (2016). Comparative genomics of biotechnologically important yeasts. Proc. Natl. Acad. Sci. U.S.a. *113*, 9882–9887.

Schorsch, C., Köhler, T., and Boles, E. (2009). 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.

Sorrells, T.R., and Johnson, A.D. (2015). Making sense of transcription networks. Cell *161*, 714–723.

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

Wolfe, K.H., and Butler, G. (2017). Evolution of Mating in the Saccharomycotina. Annu. Rev. Microbiol. *71*, 197–214.

Appendix

Methods for genetically modifying *Wickerhamomyces anomalus* and other Phaffomycetaceae species

Introduction

Answering the scientific questions that motivated the research in this thesis required the development of methods for genetically transforming the non-model yeast *Wickerhamomyces anomalus*. Fortunately, I was able to use a nourseothricin-resistance marker and transformation protocol developed for its sister species, *W. ciferrii* (Schorsch et al., 2009). These tools allowed me to construct the first reported gene deletion strains in *W. anomalus*, which let me directly test the endogenous functions of several mating-type transcriptional regulators. Using the principles previously described for *W. ciferrii* transformation, I also developed a second drug resistance marker for hygromycin selection and determined whether both markers can be used to transform other species in the Phaffomycetaceae clade. For most of the time that I worked on *W. anomalus* there were only four Phaffomycetaceae yeast with sequenced genomes. That number has recently greatly expanded, with the publication of many more yeast species genomes as part of the Y1000+ project (Shen et al., 2018). These species will likely inform future evolutionary questions that members of the Johnson Lab are interested in. In addition, many of these species are of biotechnological interest. For example, *W. anomalus* is used in winemaking and secretes toxins that can kill other fungi (Padilla et al., 2018). The methods described by Schorsch *et al.* and elaborated on here will likely aid in both evolutionary and biotechnological research by allowing for the construction of genetically altered Phaffomycetaceae species.

Results

Targeted gene deletion is possible in *W. anomalus,* despite its preference for non-homologous end joining over homologous recombination

I found that the freeze thaw transformation protocol described for *W. ciferrii* by Schorsch *et al.* works well for transforming *W. anomalus* (see Protocol below). A typical transformation with the *Nat* marker they developed (pCS.ΔLig4) yields hundreds of colonies in *W. anomalus*.

Like *W. ciferrii, W. anomalus* tends to take up selective markers by non-homologous end joining rather than homologous recombination. For example, when transforming with a vector to delete a gene of interest in *W. anomalus*, hundreds of drug-resistant colonies will grow, but the vast majority of them will not have the gene deleted. I believe that this is because they have integrated the vector randomly into the genome by non-homologous end joining.

I used a few strategies to routinely delete genes in *W. anomalus*, despite this. First, I used digested plasmid DNA as the transformation vector, rather than a PCR product. This has the advantage of generating a high yield of a clean, sequenced piece of DNA as the vector. Second, I cloned relatively long homology arms onto these plasmids. Typically, these would be around 1000 basepairs in length. At minimum, I would aim for several hundred basepairs in length. pCS.ΔLig4 was the parent plasmid for all of my *W. anomalus* gene deletion plasmids. Third, I used a relatively large amount of DNA vector in each transformation (typically 1 μg). Fourth, I screened hundreds of colonies to successfully identify deletion strains. I found that on average approximately 1/500 colonies would have the gene of interest deleted. I used colony PCR to screen transformed colonies in a 96-well format, using the same boiled sample of colony to run diagnostic PCRs to test for (1) the presence or absence of the coding-sequence of the gene I'm trying to delete, and (2) the presence or absence of one of the flanks between the Nat marker and the surrounding genome (i.e., a PCR product to bridge the junction between the inserted vector and the surrounding genome at the desired locus). Because colony PCR at this scale can be finicky, I found that always screening for both the flank and the coding-sequence was the best strategy to successfully identify true deletions.

Schorsch *et al.* additionally describe a Cre-Lox system for recycling the *Nat* marker in which they transiently express Cre off of a plasmid that is not maintained by the cell. In my hands, this has not yet worked in *W. anomalus*, and I do not know why. In addition, Schorsch et al. found

that deleting the non-homologous end joining DNA ligase gene (*DNL4* or *LIG4*) substantially increased the fraction of transformed colonies with vectors taken up by homologous recombination. I have disrupted but not deleted *LIG4* in *W. anomalus* and have never tested its effect on deleting genes by homologous recombination. I subsequently developed a second resistance marker that can be used to transform *W. anomalus* (described below), so deleting *LIG4* is likely be a useful next step for those interested in deleting genes in *W. anomalus* going forward.

***W. anomalus*-optimized hygromycin-resistance marker**

I developed a second selectable marker for *W. anomalus* using pieces of the pCS.ΔLig4 plasmid developed by Schorsch et al.. First, I cloned the P_{PDA1} promoter and T_{TEF1} terminators used to drive *Nat* expression in pCS.ΔLig4 into pUC19 (pCSB 137). I then codon-optimized the hygromycin resistance gene *hph* for expression in *W. anomalus* (see Chapter 2 Methods) and inserted it in between P_{PDA1} and T_{TEF1} (pCSB 182, isolate C). This construct successfully transformed *W. anomalus*, as well as *W. ciferrii* (Table 1). In contrast, I found that pCS.ΔLig4 successfully transformed *W. anomalus*, *W. ciferrii*, *Pichia wickerhamii*, *Cyberlindnera fabianii*, and *Cyberlindnera jadinii* (Table 1). *Pichia rhodanensis* did not yield colonies with either selectable marker.

Conclusion

The freeze thaw protocol published by Schorsch et al. reproducibly transforms multiple species in the Phaffomycetaceae clade of the Saccharomycotina yeasts. While these species tend to favor taking up selectable markers by non-homologous end joining rather than homologous recombination, targeted gene deletions are nonetheless possible by screening many transformed colonies. The development of a second selectable marker means that going forward more complex genetic manipulations will be possible in these species.

Protocol for freeze thaw transformation of *W. ciferrii*, *W. anomalus*, and other Phaffomycetaceae species (Schorsch et al., 2009)

Materials

Solution A
1 M sorbitol
10 mM bicine (pH 8.35)
3% ethylene glycol (w/v)

Solution B
40% PEG1000
200 mM bicine (pH 8.35)

Solution C
150 mM NaCl
10 mM bicine (pH 8.35)

Protocol

1. Grow yeast overnight in YEPD
2. Dilute culture back to OD_{600}=0.1-0.2
3. Grow to OD_{600}=0.6-1 (I typically let the cultures grow for four hours, aiming for cells to be in mid-log phase growth)
4. Harvest cells by centrifugation (3000 rpm for 10 minutes, 4°C)
5. Wash in 0.5 culture volumes Solution A
6. Harvest cells by centrifugation (3000 rpm for 10 minutes, 4°C)
7. Resuspend cells in 0.02 culture volumes Solution A
8. Store 200 µl aliquots at -80°C
9. For transformation, add DNA (typically at least 1µg per transformation) and 12 µl 2 mg/ml salmon sperm DNA to cells
10. Incubate at 37°C for 10 minutes taped to bottom of 200 rpm shaker (note: the original protocol calls for an 800 rpm shaker but I have found that the 200 rpm shaker that we have in the lab works fine)
11. Add 1.5 ml Solution B and mix gently (I mix by flipping the tubes up and down rather than pipetting)
12. Incubate for 1 hour at 30°C
13. Harvest by centrifugation (short spin for ~15-20 seconds on microcentrifuge, rpm at max speed)
14. Wash with 1 ml Solution C
15. Resuspend in YEPD, plate on YEPD plates, incubate at 30°C overnight (typically I will resuspend in 200-400 µl YEPD and divide the transformed cells onto at least two separate plates. I do this because most transformations will yield hundreds of colonies, most of which will have taken up the selective marker by random integration, so you need to dilute the cells sufficiently to see separate colonies)
16. The next day, replica plate onto YEPD + drug plates
17. Following replica plating, colonies should appear after 2-3 days

Species	Strain # (Candace Britton or Trevor Sorrell's logs)	Cell type (if known)	Recommended [nourseothricin] µg/ml	Recommended [hygromycin] µg/ml
W. anomalus	yCSB 5	**a** cell	50	300
W. anomalus	yCSB 8	**a**/α cell	200-400	300
W. anomalus	yCSB 342	α cell	200-400	300
W. ciferrii	yCSB 252	**a** cell	50	300
P. wickerhamii	yTS 20		50	
C. fabianii	yCSB 335		100 or 300 both worked	
C. jadinii	yTS 187		50	

Table 1. Recommended concentrations of selective drugs for transforming various Phaffomycetaceae species and strains

The Nat marker is the most broadly useful marker for transforming Phaffomycetaceae species, but different species and strains have different baseline resistances to nourseothricin. For example, while 50 µg/ml nourseothricin is sufficient to select for transformants in the *W. anomalus* **a** cell strain, the **a**/α and α strains require higher concentrations to prevent background growth. 300 µg/ml hygromycin is generally the right concentration to use to select for hygromycin-resistance, but the hygromycin-resistance marker cannot yet be used in every species. *P. wickerhamii*, *C. fabianii*, and. *C. jadinii* did not yield hygromycin-resistant colonies on plates with 300 µg/ml hygromycin but had background growth with 200 µg/ml hygromycin (that is, they grew even without a hygromycin-resistance gene transformed in).

References

Padilla, B., Gil, J., and Manzanares, P. (2018). Challenges of the Non-Conventional Yeast Wickerhamomyces anomalus in Winemaking. Fermentation *4*, 68.

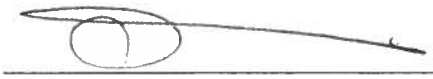
Schorsch, C., Köhler, T., and Boles, E. (2009). Knockout of the DNA ligase IV homolog gene in the sphingoid base producing yeast Pichia ciferrii significantly increases gene targeting efficiency. Curr. Genet. *55*, 381–389.

Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell *175*, 1533–1545.e20.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date March 5, 2019