

Evolution of a new form of haploid-specific gene regulation appearing in a limited clade of ascomycete yeast species

Francesca Del Frate,^{1,2} Megan E. Garber,^{1,*} Alexander D. Johnson^{1,2,*}

¹Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94102, USA

²Tetrad Graduate Program, University of California, San Francisco, San Francisco, CA 94102, USA

*Corresponding authors: Email: ajohnson@cgl.ucsf.edu (A.D.J.); megan.garber@ucsf.edu (M.E.G.)

Abstract

Over evolutionary timescales, the logic and pattern of cell-type specific gene expression can remain constant, yet the molecular mechanisms underlying such regulation can drift between alternative forms. Here, we document a new example of this principle in the regulation of the haploid-specific genes in a small clade of fungal species. For most ascomycete fungal species, transcription of these genes is repressed in the a/α cell type by a heterodimer of two homeodomain proteins, *Mata1* and *Mata2*. We show that in the species *Lachancea kluyveri*, most of the haploid-specific genes are regulated in this way, but repression of one haploid-specific gene (*GPA1*) requires, in addition to *Mata1* and *Mata2*, a third regulatory protein, *Mcm1*. Model building, based on x-ray crystal structures of the three proteins, rationalizes the requirement for all three proteins: no single pair of the proteins is optimally arranged, and we show that no single pair can bring about repression. This case study exemplifies the idea that the energy of DNA binding can be “shared out” in different ways and can result in different DNA-binding solutions across different genes—while maintaining the same overall pattern of gene expression.

Keywords: gene regulation, fungal mating, DNA binding, evolution

Introduction

Changes in transcription circuits over evolutionary timescales are a major source of phenotypic novelty. Two major sources of transcriptional plasticity have been well-documented: (1) changes in the cis-regulatory sequences of a gene, which can directly alter the pattern of expression of that gene (and indirectly affect the expression of other genes) and (2) the formation (and breaking) of cooperative interactions between different transcriptional regulators, which can directly affect the expression of many genes simultaneously (Wray 2007; Lynch and Wagner 2008; Wittkopp and Kalay 2012; Jarvela and Hinman 2015; Sorrells and Johnson 2015). Typically, the two types of changes are observed together in new circuit architectures. Both types of changes can occur without extensive pleiotropy; the former directly affects expression of only the gene in which it occurs, and the latter—because it is often due to the creation of a relatively weak protein–protein interaction in a part of the protein distinct from the DNA-binding domain—typically does not compromise the ancestral roles of the protein (Carroll 2008; Stern and Orgogozo 2008; Baker et al. 2012; Jarvela and Hinman 2015). In contrast, changes in the intrinsic DNA-binding specificity of a conserved transcription regulator over evolutionary timescales seem to occur much less frequently. In the absence of gene duplication, such changes would likely compromise the existing roles of the protein and would not be maintained.

While some evolutionary changes in transcription lead to dramatic new phenotypes, other studies indicate that the mechanisms of regulation can apparently drift between different molecular solutions while maintaining the same output (Baker

et al. 2012; Britton et al. 2020). Documenting these cases in detail provides an opportunity to understand the molecular principles behind transcription circuit plasticity. In this paper, we document and explain a clear example of this type of plasticity in the regulation of the mating genes in the ascomycete (yeast) lineage.

We concentrate on a group of genes known as the haploid-specific genes, which are expressed in the two mating cell types (a and α) but repressed in the third cell type, the a/α cell (Fig. 1a). The a/α cell is the product of the mating of an a cell and an α cell and itself is non-mating. The haploid-specific genes code for proteins required for both a and α cells to mate; for example, three genes code for the components of the trimeric G protein needed for pheromone signaling (Herskowitz 1989). Their repression in the a/α cell therefore makes logical sense as their gene products are not needed, and could even be detrimental, in this cell type.

In many ascomycetes, the haploid-specific genes are repressed directly by a heterodimer of two homeodomain proteins, *Mata1* and *Mata2* (Strathern et al. 1981; Tsong et al. 2003; Galgoczy et al. 2004). Again, this logic makes conceptual sense: *Mata1* is made by a cells and *Mata2* by α cells; only when the two proteins are synthesized together in a/α cells (the result of mating) does the heterodimer form and repress the haploid specific genes.

Although direct repression by the *Mata1*–*Mata2* heterodimer is highly logical and greatly appealing in its simplicity, there are exceptions to this mechanism. In *Kluyveromyces lactis*, the repression of the haploid-specific genes is indirect: the *Mata1*–*Mata2* heterodimer represses an activator of the haploid-specific genes but does not bind these genes directly (Booth et al. 2010). And in *Wickerhamomyces*

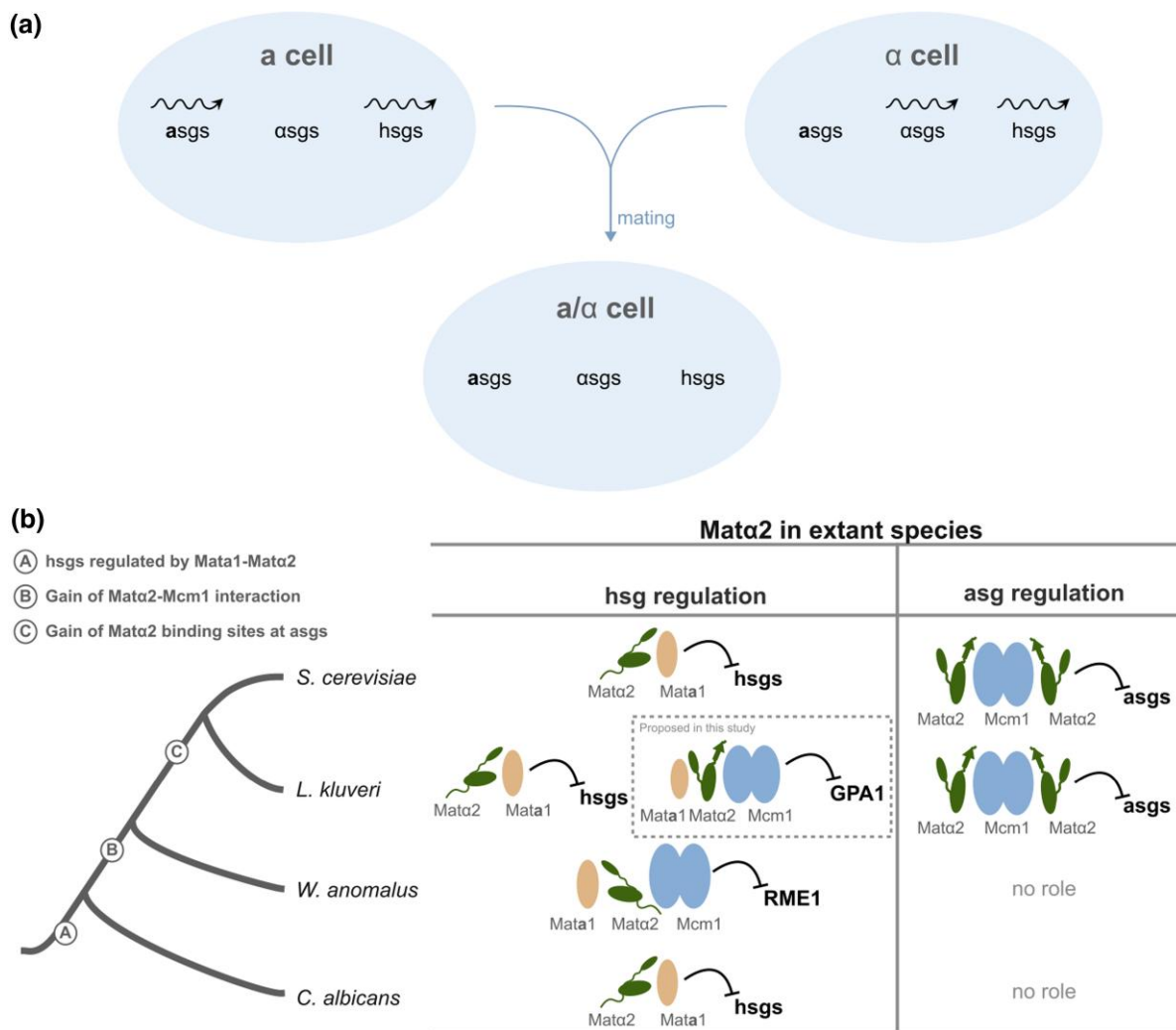


Fig. 1. Regulation of cell type in budding yeast. a) Three cell types in budding yeast; **a** and α cell types express the **a**-specific and the α -specific genes, respectively, and both **a** and α cell types express the haploid specific genes. When **a** and α cells mate, the resulting **a/α** cell does not express these genes. b) Regulation of **a**-specific genes and haploid-specific genes by Mata2 and its binding partners Mata1 and Mcm1 in yeast species. Haploid-specific genes are directly regulated by the Mata1-Mata2 heterodimer in *S. cerevisiae* and *C. albicans* and likely constitute the ancestral form of regulation (indicated by circled A on the figure). On the branch leading to *W. anomalus* and *S. cerevisiae*, a protein-protein interaction gained between Mata2 and Mcm1 (circled B) allowed for the addition of Mcm1 to haploid specific gene regulation in *W. anomalus* and a gain of Mata2-Mcm1 repression at the **a**-specific genes in the lineage leading to *S. cerevisiae* (circled C). As proposed in this study, regulation of haploid specific genes in GPA1 in *L. kluyveri* is by tripartite Mata1-Mata2-Mcm1. This mechanism requires both the Mata2-Mata1 interactions (circled A) and the Mata2-Mcm1 interactions (circled B) to have previously formed and, as described in the text, there is strong evidence for this sequence of events. The other conserved haploid-specific genes in *L. kluyveri* are regulated by Mata1-Mata2 as in *S. cerevisiae* and *C. albicans*.

anomalus, the Mata1-Mata2 heterodimer requires a third protein, Mcm1, to repress at least one of the haploid-specific genes (Britton et al. 2020).

In this paper, we investigated regulation of the haploid-specific genes in *Lachancea kluyveri*, a species that branched from *S. cerevisiae* well after the *S. cerevisiae*-*W. anomalus* branchpoint (Fig. 1b). We were drawn to this species because bioinformatic analyses indicated that one of the haploid-specific genes (GPA1, which codes for the alpha subunit of the trimeric G protein) appeared to lack a conventional Mata1-Mata2 heterodimer binding site, whereas other haploid genes in this species (and in many other species) clearly displayed this signature motif (Booth et al. 2010). In this paper, we show that the *L. kluyveri* GPA1 is not regulated in the conventional, deeply-conserved manner but is repressed in the **a/α** cell by three proteins working together, Mata1, Mata2, and Mcm1. In this three-part regulatory complex, we show that any

pair of proteins is not sufficient to bring about repression due to the non-optimal arrangement of their cis-regulatory sequences, resulting in the requirement for all three proteins. Mcm1 is produced in all three cell types, so the logic of regulation is preserved: repression occurs only in the **a/α** cell type, despite the idiosyncratic arrangement of proteins on DNA. We discuss possible evolutionary pathways that could have led to this non-canonical mechanism of regulation.

Methods

Reporter constructs

Reporter constructs were made using TS185, a plasmid containing a hygromycin resistance cassette previously used to stably integrate a GFP transcriptional reporter at the URA3 locus in *L. kluyveri* (Sorrells et al. 2015). The plasmid included restriction sites for

Age1 and BsiWI allowing for insertion of control sequences to test their effect on gene expression. Custom Geneblocks (Integrated DNA Technologies, gBlocks Gene Fragments) were designed with 500 base pairs upstream of the GPA1 transcriptional start site and mutant versions as described in Fig. 3 (Supplementary File 1). These include a wild type GPA1 upstream sequence, and GPA1 upstream sequences where the motifs for Mata1, Mata2, and Mcm1 are individually scrambled. Scrambled sites contained as many changes in the site as possible, while maintaining overall GC content. A sequence with all three of the motifs scrambled was also constructed, and a gg→cc point mutation in conserved residues of the putative Mcm1 site was also included.

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

Strain construction

Construct plasmid DNA was linearized by NotI-HF and EcoRV-HF digest to prepare for transformation into yeast. *L. kluyveri* α cells were transformed by electroporation (Faber et al. 1994; Gojkovic et al. 2000) with the following modification: after electroporation cells were collected in (Faber et al. 1994; Gojkovic et al. 2000) 1 mL yeast extract–peptone–dextrose (YPED) and plated to non-selective YPED plates to develop into a lawn overnight at 30°C. The next day, cells were replica plated to 400 µg/mL Hygromycin YPED plates and grown overnight at 30°C. Colonies that arose after 24 hours were patched to SC-ura and YPED 5-FOA plates (0.8 mg/mL 5-FOA in YPD agar) for a second round of selection. Isolates that were Ura- and Hygromycin resistant were grown overnight in 2 mL of YPED for DNA extraction (Hoffman and Winston 1987). Cells were spun down, resuspended in 200 µL lysis buffer (2% v/v Triton X-100 1% v/v SDS 100 mM NaCl 10 mM Tris-Cl pH 8.0 1 mM EDTA pH 8.0), 200 µL phenol chloroform pH8 (Fisher Scientific 68-051-00ML). Two hundred microliter 0.5 mm glass beads (BioSpec Products 11079105) were added, and samples were lysed for 5 min in a benchtop vortexer. Lysed cells were then centrifuged at 18407 rcf for 2 min. 200 µL of the supernatant was taken out and precipitated in 1 mL of ethanol. Using the isolated DNA, strain construction was confirmed by PCR of upstream and downstream flanks of the insertion at the URA3 locus and lack of the URA3 open-reading frame. Three independent transformants were validated for each construct.

The three isolates of each validated α strain were mated to *L. kluyveri* **a** cells (LB76) by mixing and spotting roughly equal amounts of cells from fresh colonies of each cell type onto a fresh YPED plate. After allowing the cells to mate for 3 hours at 30°C, the spotted cells were plated for single colonies on YPED. Single colonies were patched onto SC-ura plates, YPED 5-FOA plates, and YPED Hyg plates. Isolates that were Ura+ and Hyg+ were validated as diploid **a/a** cells by PCR checks for both the MAT α and MAT α locus using extracted gDNA. One **a/a** strain per α isolate was validated and saved, so that each independent transformant would have a matched **a/a** strain.

The tagged Mata2 **a/a** strain, FDy18, used in the chromatin immunoprecipitation experiment (see below) was generated from an existing strain used in a previous study (yLB96) which had a c-terminal 13× Myc-tag on the endogenous Mata2 in an α cell (Baker et al. 2012). This strain was mated with a naïve strain (LBy76) of the **a** cell type as described above to generate the c-terminally Myc-tagged Mata2 strain in the **a/a**. The untagged strain, FDy22, was generated from mating yLB76 and yLB77, the prototrophic **a**

and α strains. All strains used in this study are listed in Supplementary Table 1.

RNA-Seq

Cultures were inoculated from single colonies and grown overnight in YPED at 30°C, diluted back to an OD600 of 0.15 in the morning and harvested at an OD600 of 0.6–0.9 as is described (Nocedal et al. 2017). Three replicates of yLB76 (**a** cell) and yLB77 (α cell) were generated from individual single colonies grown from the same streak. Three replicates from FDy22 (**a/a** cell) were from three independently mated isolates. RNA was extracted using the RiboPure RNA purification kit (ThermoFisher AM1924). Total RNA quality was verified on an Agilent TapeStation. Total RNA was poly-A selected with the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB E7490S). cDNA synthesis and library preparation was carried out with the NEBNext Ultra II Directional RNA Library Prep kit for Illumina (NEB E7760L). Quality and concentration of libraries were determined with an Agilent TapeStation. Libraries were pooled in equimolar amounts and sequenced using single end 65 base pair reads on an Illumina HiSeq 4000.

RNA-Seq analysis

Quality of sequencing reads was determined using FastQC (Andrews 2010). Filtering based on quality and trimming of reads was carried out using FastP (Chen et al. 2018). (Consortium et al. 2009; Shen et al. 2018) Trimmed reads were aligned to the *Lachancea kluyveri* NRRL Y-12651 (Consortium et al. 2009; Shen et al. 2018) reference genome using STAR (Dobin et al. 2013). A table with counts assigned to genes was generated from the alignments using Rsubread (Chisanga et al. 2022). This count table was then used to determine differentially expressed genes using DESeq2 (Love et al. 2014). DESeq2 was run with default parameters, resulting in a list of genes that were differentially expressed in yLB76 (**a** cell) or yLB77 (α cell) when compared to FDy22 (**a/a** cell) (Supplementary File 2–3). Genes with an adjusted $P < 0.1$ in the **a** cells vs **a/a** cells and the α cells vs **a/a** cells were plotted against each other in GraphPad Prism. Genes up-regulated greater than a \log_2 (fold-change) of 2 and an adjusted $P < 0.1$ in both differential expression comparisons (**a** vs **a/a** and α vs **a/a**) were interpreted as significantly enriched for expression in the haploid cell types.

ChIPseq

Chromatin immunoprecipitation was carried out as previously described (Sorrells et al. 2015; Nocedal et al. 2017) using a myc-tagged Mata2 **a/a** strain (FDy18), a myc-tagged Mata2 **a** strain (yLB77), a myc-tagged Mata2 α strain (yLB96), and an untagged strain (FDy22), with the following modifications. Cells were lysed by bead beating using 0.5 mm Zirconia beads in the Omni Bead Ruptor 12 using three 90 s cycles, alternated with 90 s of cooling samples on ice. The isolated chromatin was sheared by sonication in a Diagenode Bioruptor Pico cycling 30 s on, 30 s off, for 25 min. Immunoprecipitation was carried out using Invitrogen Anti-c-Myc Monoclonal (9E10.3) Antibody (ThermoFisher AHO0062). Sepharose Protein G beads were replaced with 30 µL of Dynabeads (ThermoFisher 10004D).

Libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB E7645L). Individual library quality and concentration were assessed by Agilent TapeStation. Libraries were pooled in equimolar amounts for single end 65 base pair reads using an Illumina HiSeq4000 at the UCSF Center for Advanced Technology.

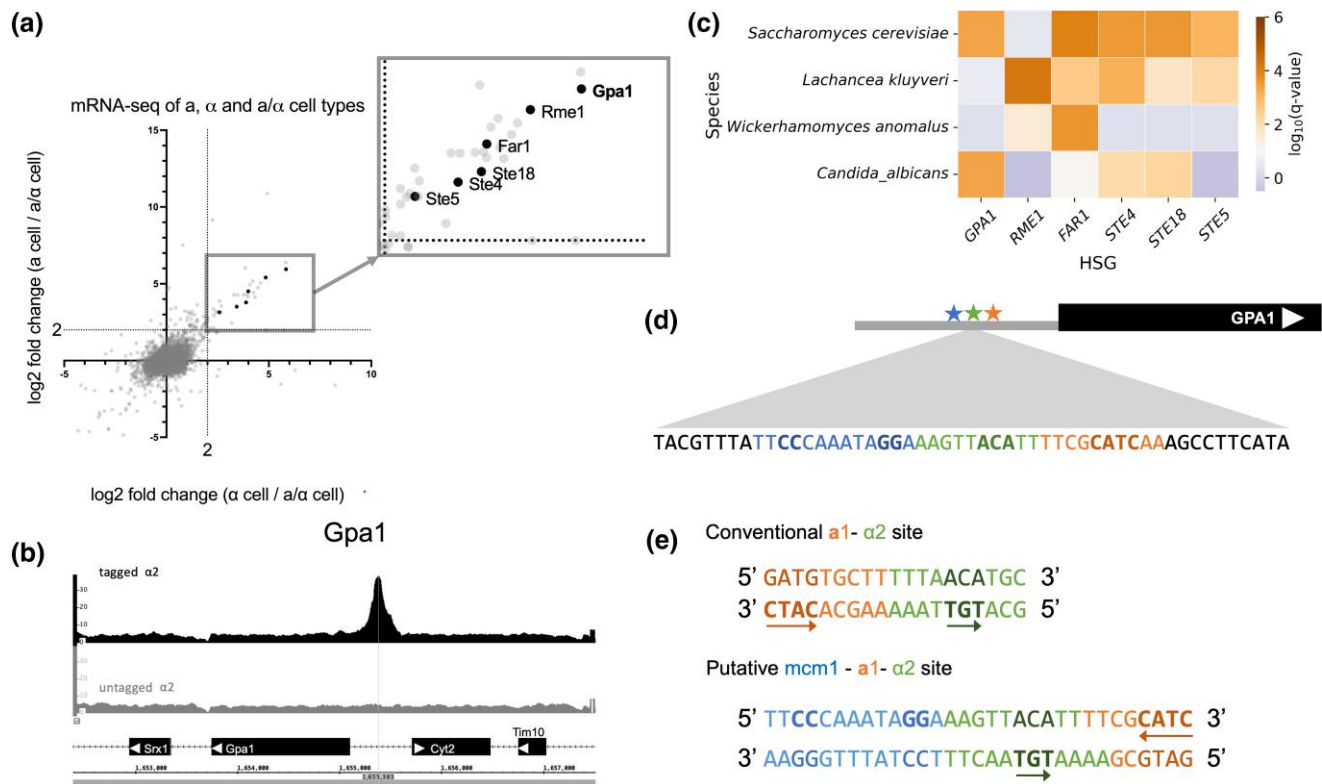


Fig. 2. Haploid-specific regulation in *L. kluyveri*. a) RNAseq results from an **a** cell, α cell, and **a/a** cell in *L. kluyveri*. Genes expressed at higher levels in both the **a** and α relative to the **a/a** ($\log_2(\text{fold change}) > 2$ and $P < 0.1$) defined as haploid specific genes in *L. kluyveri*. Inset panel shows the expression pattern of the conserved haploid-specific genes GPA1, RME1, FAR1, STE4, STE18, and STE5. b) Chromatin immunoprecipitation of a C-terminal myc-tagged Mata2 in an α cell shows significant enrichment at the promoter of GPA1. Results with the tagged strain are shown in black compared to the matched untagged strain in grey. c) Bioinformatic search for the conventional Mata1–Mata2 motif in upstream regions of GPA1, RME1, FAR1, STE4, STE18, and STE5 in *S. cerevisiae*, *W. anomalus*, *L. kluyveri*, and *C. albicans*. *L. kluyveri* GPA1 upstream region lacks a high scoring Mata1–Mata2 site, while the other five hsgs all have high scoring Mata1–Mata2 sites. Site score is \log_{10} of the reported q -value. d) Schematic of predicted Mcm1, Mata2, and Mata1 sites found in the *L. kluyveri* GPA1 control region. The Mcm1 motif is indicated in blue, Mata2 motif in green, and Mata1 motif in orange. Bases identical to those in the corresponding *S. cerevisiae* motifs are bolded. e) Conventional **a1**–**a2** site found upstream of *S. cerevisiae* GPA1 (top) and the pMcm1–**a1**–**a2** site found upstream of *L. kluyveri* GPA1 (bottom). The Mcm1 motif is indicated in blue, Mata2 motif in green, and Mata1 motif in orange. Bases identical to those in the corresponding *S. cerevisiae* motifs are bolded. Arrows below the bases, which represent the points of contact for **a1** or **a2** on double stranded DNA, are drawn 3'→5' to indicate that the relative orientation of the **a1** and **a2** sites is “backward” in the three-part sequence when compared to the orientation in the conventional sequence.

ChIPseq analysis

Reads were trimmed and aligned as described above for the RNAseq. BAM files were processed using DeepTools (Ramírez et al. 2016) and Samtools (Danecek et al. 2021) and uploaded to the Integrated Genomics Viewer (Robinson et al. 2011) and the Integrated Genome Browser (Freese et al. 2016) for visual inspection of data. Visual inspection included comparison to data tracks for the myc-Mata2 tagged **a** strain (yLB77) or the myc-Mata2 tagged α strain (yLB96). Peaks found in either tagged **a** or α strain were removed from consideration, resulting in a set of peaks unique to the **a/a** cell. Data was processed with MACS2 (Gaspar 2018) with default settings except for those regarding duplicates (Gaspar 2018). Instead of removing all duplicate reads, the MACS2 function for keeping biologically relevant duplicates was used, an adjustment recommended for transcription factors with few targets in samples with high read depth.

qRT-PCR of transcription reporter

Single colonies from three independent genetic isolates of each of the 12 reporter strains (FDy27, FDy28, FDy30, FDy31, FDy32, FDy33, FDy34, FDy35, FDy36, FDy37, FDy38, FDy39, see above for

strain construction) were inoculated in YPED and grown overnight at 30°C, then diluted back to an OD600 of 0.1 the following morning. Outgrowth cultures were harvested between an OD600 of 0.7–0.9, flash frozen in liquid nitrogen and stored at –80°C. RNA was extracted from frozen cell pellets using the MasterPure-Yeast-RNA Extraction Kit (Lucigen MasterPure Yeast RNA Purification Kit MPY03100) with one modification. After the isopropanol precipitation, RNA was treated with TURBO DNA-free kit (ThermoFisher AM1907). The isolated RNA was reverse transcribed with the Superscript III Reverse transcriptase kit (ThermoFisher 18080044) using 250 ng of random primers. qRT-PCR was carried out using 2× iTaq Universal SYBR Green Supermix (Bio-Rad 1725124) and a Bio-rad CFX96 Real-Time PCR to measure cDNA amplification. qPCR probes against GFP were designed using the NCBI Primer-BLAST tool (Ye et al. 2012). Previously verified probes for ACT1, a housekeeping gene between cell types in *L. kluyveri*, were used in this study (Booth et al. 2010). Ct values were calculated with CFX Maestro software (Bio-rad 12004128). GFP expression was normalized to ACT1 and to overall expression of all samples. Expression from constructs with the various site mutants was compared to expression from the construct with the wild type sequence to calculate fold repression.

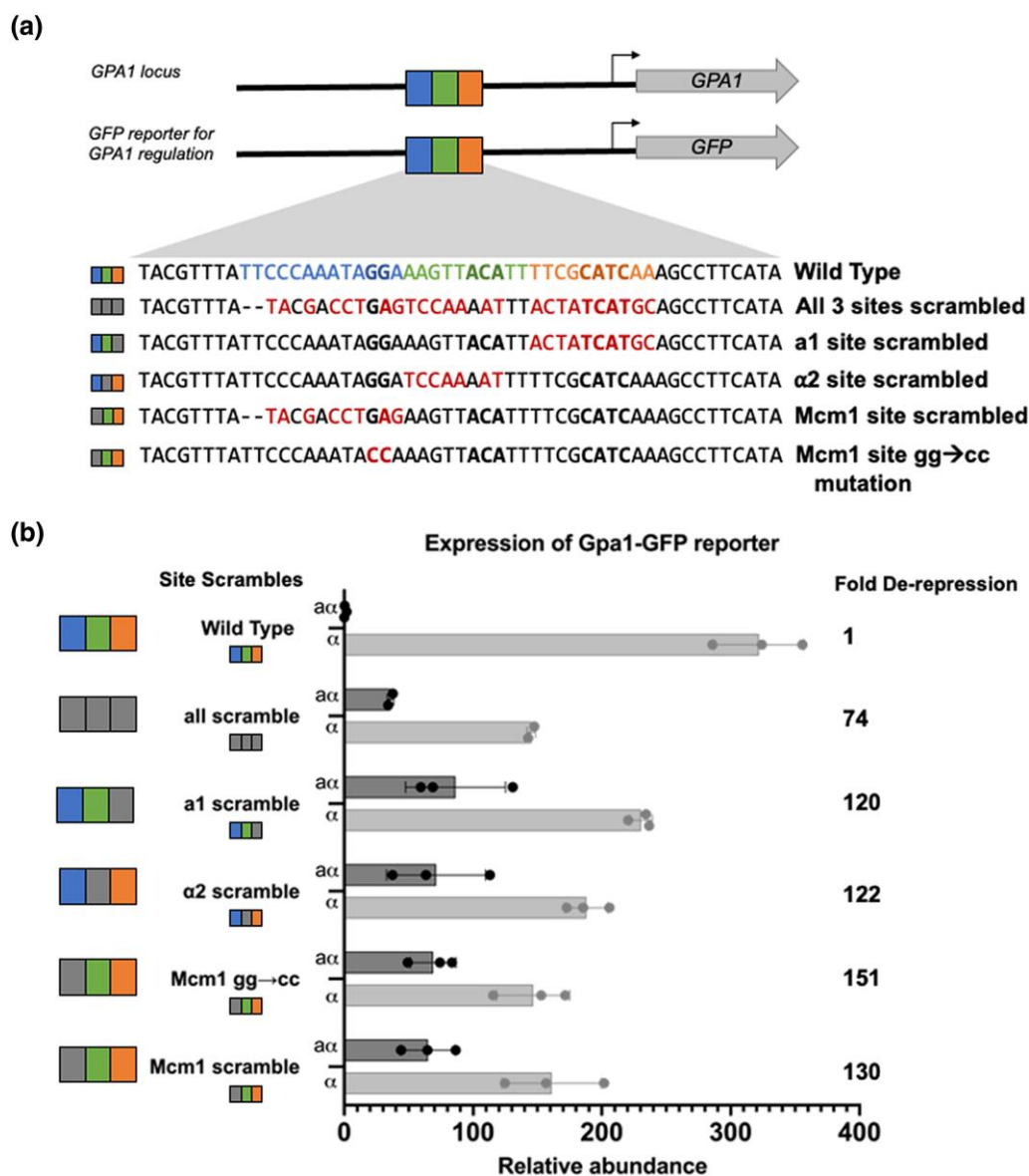


Fig. 3. Tripartite regulation of *GPA1*. a) In these constructs, the upstream region of *GPA1* drives expression of *GFP*. Wild type (top line) and mutated *GPA1* control sequences are shown. In the wild type sequences, the *Mcm1* motif is indicated in blue, *Mata2* motif in green, and *Mata1* motif in orange. Mutated bases in the *GPA1* control sequences are indicated in red. Bases identical to those in the corresponding *S. cerevisiae* motifs are bolded. b) Expression of *GFP* reporter transcript in *L. kluyveri* in α and a/α cells. Expression was measured by qPCR with probes to *GFP* transcript normalized to *ACT1*. The results show the average of three independent genetic isolates for constructs 1 and 3–6, and two independent genetic isolates for construct 2. Each genetic isolate was measured twice in independent biological replicates with standard deviations shown. The right-most column (fold derepression) are the ratios of mutant construct expressions compared to the wild type construct, all in a/α cells. These numbers are scaled to different levels of expression in the α cell.

Modeling *Mcm1*, *Mata1*, and *Mata2* as a complex bound to DNA

Using ChimeraX (Goddard et al. 2018; Pettersen et al. 2021), the structure for *Mcm1* and *Mata2* bound to DNA (1YRN) (Li et al. 1995) and the structure *Mata1* and *Mata2* bound to DNA (1MNM) (Tan and Richmond 1998), retrieved from the Protein Data Bank (Berman et al. 2000), were overlaid to reflect the spacing and orientation of the three sites as observed in the tripartite site. Using the “move model” and “rotate model tools”, the *Mata1*–*Mata2* structure (1YRN) was rotated and moved relative to 1MNM (the structure of *Mcm1* and *Mata2*), arranging the two such that the DNA in the two structures aligned. Conserved DNA residues with key protein contacts for *Mata2* were six base pairs away from the conserved DNA residues with key protein

contacts for *Mata1*. Proteins were docked to their binding sites on the DNA; this resulted in the arrangement of *Mcm1*–*Mata2*–*Mata1* proteins along the overlapped DNA, with *Mata1* on the opposite site of *Mata2* (compared with 1YRN) (Supplementary File 4).

Measuring the distance spanned by *Mata2* C-terminal linker

Using the “tape” tool in ChimeraX (Goddard et al. 2018; Pettersen et al. 2021), the distance was measured from the C-terminal residue of the third α -helix (α_3) in *Mata2* (1MNM #2/C Thr 189) and the N-terminal residue of the fourth α -helix (α_4) *Mata2* that interacts with *Mata1* (1YRN #1/B Pro194). This distance, calculated by drawing a straight line between the two residues, is the minimum distance that the flexible linker region of *Mata2* needs to span. To

compare the distance spanned by the linker in the existing structure (1YRN) to the distance spanned by the linker in the model, the distance between the corresponding residues on Mata2 on the 1YRN structure (1 YRN #1/B Thr 191 and 1 YRN #1/B Pro194) were also measured on 1YRN using the same approach. The extended length of the amino acid linker was calculated by multiplying the number of amino acids by the average length per amino acid as empirically determined (Ainavarapu et al. 2007).

Generating position specific scoring matrices of DNA binding motifs

Mata1–Mata2 motif: The 1000 bp upstream regions of the 12 haploid specific genes in *S. cerevisiae* were extracted using the SGD Sequence Resources Tool (Cherry et al. 2012; Engel et al. 2013). These were input into MEME with default settings to generate a Mata1–Mata2 motif (Supplementary File 5) (Bailey et al. 2015).

Mata1–Mata2–Mcm1 motif: Sequences in the *L. kluyveri* genome in which ChIP signal was enriched were extracted using the Integrated Genomics Viewer, and inputted into MEME to generate de novo motifs (Bailey et al. 2015; Robinson et al. 2011). This sequence was then refined using the *S. cerevisiae* motifs to generate a synthetic position specific weight matrix for the tripartite Mata1–Mata2–Mcm1 site in *L. kluyveri*. This consisted of flipping the orientation of the Mata2 motif relative to the Mata1 motif—so that the relative orientation and spacing of the two motifs matches that of the tripartite site upstream of GPA1 in *L. kluyveri*—and adding an *S. cerevisiae* Mcm1 motif downloaded from the Jaspas database (Castro-Mondragon et al. 2021), setting the spacing to match the tripartite site in *L. kluyveri* (Supplementary File 6).

Bioinformatics search for binding sites upstream of haploid specific genes

We identified orthologs across budding yeasts for the haploid specific genes, GPA1, RME1, FAR1, STE4, STE18, and STE5, by mining data made available by the 1000 yeast genomes project (Shen et al. 2018) (Supplementary File 7–12). For each identified ortholog, we used its coordinates and direction (positive or negative strand) to append an entry of 1000 bp upstream of the gene of interest in its respective genome into a fasta file named after the gene of interest (Supplementary File 13–18). We applied FIMO (Grant et al. 2011) to search for the respective motifs specified in the text, figures, and visualized in Fig. 5 with default options and a statistical threshold (P -values) of 1×10^{-2} . The data were visualized by concatenating the highest scoring $-\log_{10}(q\text{-value})$, from each independent FIMO search. The resultant high scoring hits were visualized in a heatmap, where orthologs are sorted by their phylogenetic orientation as previously determined (Shen et al. 2018).

Results

Regulation of the haploid specific genes in *Lachancea kluyveri*

We first identified the haploid-specific genes in *L. kluyveri* by comparing gene expression across the three cell types: **a**, **α** , and **a/α** (Fig. 1a). Haploid-specific genes are defined here as genes that are expressed in **a** and **α** cells but not in **a/α** cells. By RNA-seq, we identified 30 haploid-specific genes, including those encoding the three subunits of the trimeric G protein that mediates pheromone response (GPA1, STE4, STE18), the cyclin dependent kinase inhibitor (FAR1) that triggers cell cycle arrest as part of the mating response and RME1, a transcription regulator with a variety of

functions (Fig. 2a, Supplementary Table 2) (Herskowitz 1989). These five genes are haploid specific genes in many other fungal species, indicating a deeply conserved expression pattern, and these are the genes we concentrate on for the remainder of the paper (Herskowitz 1989; Booth et al. 2010; Britton et al. 2020).

As discussed in the introduction, the haploid-specific genes in many species are repressed by direct binding of the Mata1–Mata2 heterodimer, both subunits of which are synthesized only in the **a/α** cell (Strathern et al. 1981; Galgoczy et al. 2004). To test whether this is the case in *L. kluyveri*, we performed a chromatin immunoprecipitation using tagged Mata2 in the **a/α** cell (Fig. 2b, Supplementary Fig. 1, Supplementary Table 3). We identified seven high-confidence peaks including those spanning the upstream regions of GPA1, STE4, STE18, FAR1, and RME1 (Fig. 2b and Supplementary Fig. 1, Supplementary Table 3). A bioinformatic search found conventional Mata1–Mata2 motifs upstream of only four of these genes (Fig. 2c; Supplementary Fig. 2a). The exception was GPA1 where the motif appeared to be missing—even though GPA1 exhibited clear haploid specific gene expression and an obvious Mata2 ChIP signal (Fig. 2, a and b, c). This apparent paradox led us to manually examine the DNA sequence under the Mata2 ChIP peak. We identified a Mata2 DNA sequence motif and a Mata1 motif, but the orientation of the Mata2 motif was “backwards” relative to the Mata1 motif, and the spacing between the two motifs was three base pairs shorter than that of the conventional heterodimer site (Fig. 2e, Supplementary Fig. 2). These differences explain the failure of a position-weighted motif searching algorithm (based on the conserved heterodimer site) to highlight this site (Fig. 2c). We also identified an adjacent, two-fold symmetric motif for Mcm1, a protein known to interact with Mata2 for a different role in the cell, repression of the **a**-specific genes in **α** cells (Fig. 2d; Supplementary Fig. 2c). Thus, it appeared as though three proteins (and three sequence motifs) were required to repress GPA1 in *L. kluyveri*, while the other haploid-specific genes in this species contained all the hallmarks of regulation by the conventional Mata1–Mata2 heterodimer.

Testing the three-site hypothesis

To test this model of tripartite regulation of GPA1 in *L. kluyveri*, we mutated each of the three sites and measured the effects on repression. To avoid disturbing regulation of the endogenous GPA1 gene (which could have consequences such as cell-cycle arrest), we created reporter constructs with the sequence upstream of GPA1 driving the expression of GFP, which we integrated into the genome (Fig. 3a). To capture the full dynamic range of regulation, we used qPCR, rather than fluorescence (which has a relatively high background), to directly measure transcript levels. Mutations to the three-part site included independently scrambling each of the three motifs and scrambling all three sites at once. In addition, we constructed a double point mutation in the Mcm1 motif, a change known to destroy binding of Mcm1 to DNA. We know from the expression data in the three cell types that both Mata1 and Mata2 proteins are required for repression of GPA1 (Fig. 2a, Supplementary Table 2). We could not test Mcm1 in a similar way because it is essential; however, the double point mutation in the Mcm1 binding motif is more specific to Mcm1 than is a scrambled site and thus links the protein to the site.

All of these manipulations disrupted repression, showing that all three sites are needed for proper regulation (Fig. 3b). In contrast, expression in the **α** cell is relatively unaffected, so it is unlikely that the tripartite motif plays a major role in the

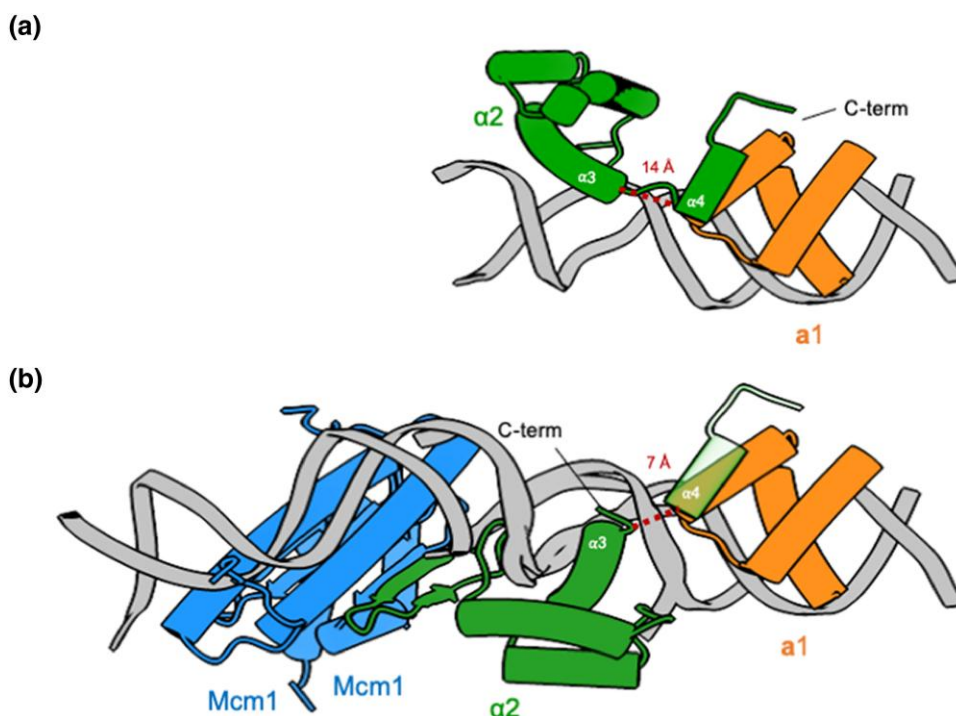


Fig. 4. Proposed arrangement of Mcm1, Mata2, and Mata1 on the *L. kluyveri* GPA control region. a) Crystal structure of the DNA-binding domains of *S. cerevisiae* Mata1–Mata2 bound to DNA as previously determined by x-ray crystallography, with Mata1 in orange and Mata2 in green. The third helix of the DNA binding domain of Mata2 is labeled $\alpha 3$ and the C-terminal helix that forms upon interaction with Mata1 is labeled $\alpha 4$. The minimum distance between $\alpha 3$ and $\alpha 4$ (indicated by the red dashed line) is ~ 14 Å. This distance is spanned by an extended linker region of four amino acids. b) Crystal structure of Mcm1–Mata2 bound to DNA in *S. cerevisiae* with the proposed arrangement of Mata2 in the “backward” orientation with respect to Mata1. Mata1 is depicted in orange, Mata2 in green, and Mcm1 in blue. As in (a), the third helix of the DNA binding domain of Mata2 is labeled $\alpha 3$. One question addressed by this modeling is whether the $\alpha 4$ helix from Mata2 (transparent green) can reach the interaction surface of Mata1 as it does in the Mata1–Mata2 structure (a). For Mata2 to interact with Mata1 in this model, the flexible linker region of 4 amino acids on Mata2 must span ~ 7 Å. The 4-amino acid linker measures ~ 15 – 16 Å when extended [as depicted in (a)], and the known structures are therefore consistent with the model.

activation of GPA1; rather, it seems to be dedicated solely to repressing the gene in **a/a** cells.

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

Having demonstrated that all three motifs are needed for repression of GPA1 in **a/a** cells, we next considered how the three proteins might be arranged on this control region and whether this arrangement provided insights into this mode of regulation. As discussed above, the motif corresponding to Mata2 lies between the motifs corresponding to Mata1 and Mcm1 (Fig. 2d). Superposition of the preferred motif for each protein onto the three-part site therefore strongly indicated that Mata2 was located between Mcm1 and Mata1 (Fig. 4, Supplementary Fig. 2). When positioned on DNA using matches with their individual motifs, the spacing between Mcm1 and Mata2 is exactly the same as it is when the two proteins interact to repress the **a**-specific genes (Fig. 2d, Supplementary Fig. 2). It therefore seems very likely that the same arrangement of Mcm1 and Mata2 (which allows a favorable protein–protein interaction between the two proteins) occurs on both the **a**-specific genes (observed in many species) and on the haploid-specific gene GPA1 in *L. kluyveri*. Regarding Mata1, inspection of the sequence showed a strong match to its motif. However, when all three proteins are placed on DNA to match their motifs, Mata2 is positioned “correctly” to interact with Mcm1, but is forced into a “backwards” orientation relative to Mata1, when compared with the conventional, heterodimer arrangement (Supplementary Fig. 2). This change in orientation is

accompanied by a change in the distance between the Mata2 motif and the Mata1 motif; it is shorter by three base pairs in the GPA1 site than in the conventional motif (Fig. 2d; Supplementary Fig. 2).

To investigate this model further (in particular to determine if there are any steric clashes), we used the solved crystal structures of *S. cerevisiae* Mcm1–Mata2 bound to DNA and *S. cerevisiae* Mata1–Mata2 bound to DNA to model the tripartite complex on DNA (Wolberger et al. 1991; Li et al. 1995) (Fig. 4). All three proteins are spatially well accommodated on their preferred motif, with the only remaining question being how Mata1 and Mata2 might interact on the GPA1 control region, given the differences in orientation and spacing from the conventional heterodimer site. In *S. cerevisiae*, Mata2 interacts with Mata1 through a short alpha helix at the end of a flexible region; the helix forms only when the two proteins interact. Comparison of the *S. cerevisiae* Mata2–Mata1 heterodimer structure to Mata1 and Mata2 as positioned on the GPA1 site in the tripartite complex suggests that the short alpha helix of Mata2 can easily reach the same position of Mata1, indicating that, despite the spacing and orientation differences, the two proteins may interact with each other through the same interface (Fig. 4). However, there must be a severe energetic cost to this altered, non-optimal configuration: when the GPA1 Mata1–Mata2 site is tested alone (that is, when the Mcm1 motif is mutated) repression by Mata1 and Mata2 is deficient (Fig. 3b). In contrast, the conventional Mata1–Mata2 motifs at the other haploid-specific genes in *L. kluyveri* (where the arrangement of sites is optimal) do not contain an Mcm1 motif.

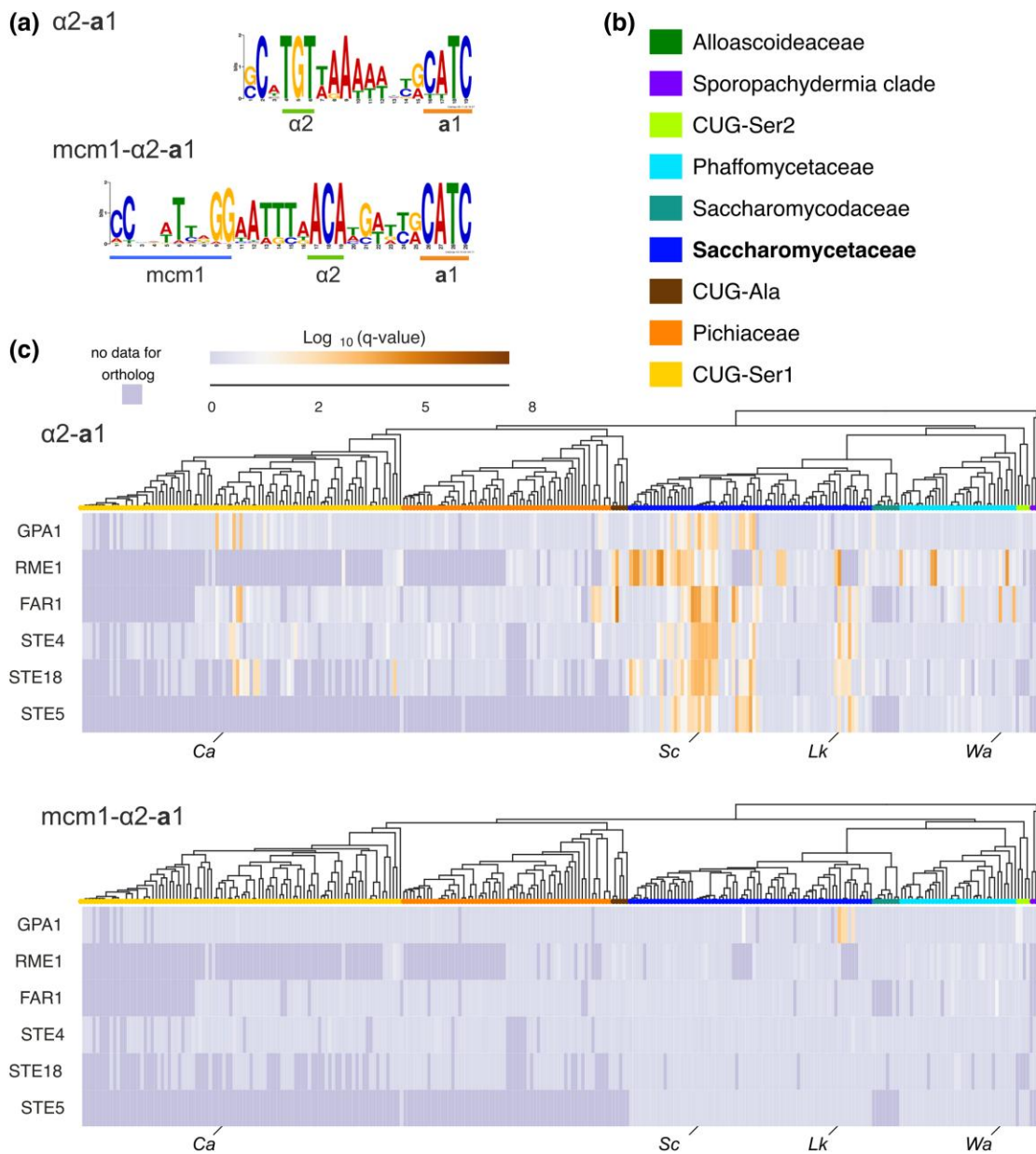


Fig. 5. Distribution of haploid-specific gene control motifs. a) Position-specific weight matrices used in the bioinformatics motif searches of the conserved haploid-specific genes across species. Top, the motif for the Mata1–Mata2 derived from *S. cerevisiae*; bottom, motif generated and refined by combining the position specific weight matrix of each individual site in the orientation of the Mcm1–a1– $\alpha 2$ site found upstream of *L. kluyveri* GPA1. b) Legend for colors used to indicate different clades across the yeast tree. c) Motifs were used to search upstream regions of orthologs of the haploid specific genes (GPA1, RME1, FAR1, STE4, STE18, and STE5) across yeast species. The best possible match to the site is given a q -value, a color is assigned based on that q -value, with pale lavender to white indicating low significance, and dark orange-brown indicating high significance. From top to bottom of each panel are gene names, each row shows the scores for the orthologs of that gene across the species. A phylogenetic tree (Shen et al. 2018) indicates the relatedness of the various species used for this study. Clades are indicated by colors in the order shown in (b).

Discussion

In this paper, we investigate a regulatory system that is deeply conserved in the fungal lineage, namely, repression of the haploid-specific genes by a heterodimer composed of one subunit of the homeodomain protein Mata1 and one subunit of the homeodomain protein Mata2. This is one of the simplest forms of regulation imaginable: One of the subunits (Mata1) is made in **a** cells and the other (Mata2) is made in α cells; only in **a**/ α cells, which arise from mating (by cell fusion) between **a** and α cells, are both halves of the heterodimer made in the same cell and the haploid genes repressed. The haploid-specific genes include those that

are needed for both **a** and α cells to mate; for example, they encode the components of the trimeric G protein needed for both cell types to respond to mating pheromones.

This simple regulatory scheme is found throughout the ascomycete lineage (Fig. 1b). This lineage represents approximately the same degree of divergence as that between humans and sponge; therefore, the conventional heterodimer regulatory scheme is widely distributed (Taylor and Berbee 2006; Shen et al. 2018).

Despite its conservation and appealing simplicity, we show that this regulatory scheme has a notable variation observed in *L. kluyveri*. In this species, most of the haploid-specific genes are

regulated in the conventional manner, but GPA1 has a novel regulatory scheme that differs in several important ways from the conserved scheme. Specifically, we show that repression of GPA1 in **a/a** cells of *L. kluyveri* requires binding of Mata1, Mata2, and a third protein Mcm1. When positioned on DNA using motif analysis and prior crystal structures, it becomes clear why no single pair of proteins suffice to bring about repression of GPA1, even though two proteins are sufficient in other contexts (Fig. 4). As shown in Fig. 1b, Mcm1 and Mata2 are positioned on GPA1 DNA exactly as they are when the two proteins carry out a different regulatory function, repression of the **a**-specific genes. This positioning results in a favorable contact between the two proteins, resulting in their cooperative binding to DNA. Despite this favorable orientation, Mcm1 and Mata2 cannot repress GPA1 alone—Mata1 is also required (Fig. 3). The reason for the failure of Mcm1 and Mata2 to work alone on GPA1 is obvious from prior work: repression of the **a**-specific genes by these two proteins requires two binding sites for Mata2, one on each side of Mcm1. If one site is experimentally mutated, repression of **a**-specific genes is destroyed (Baker et al. 2012). Thus, the configuration of Mcm1 and Mata2 on GPA1 resembles a mutant **a**-specific gene regulatory site and, based on prior work, would not be expected to function, a prediction borne out by direct experiment (Fig. 3b).

The Mata1–Mata2 pair is also insufficient to repress GPA1, and the likely reason for this is also clear. The orientation of the Mata2 subunit is “backwards” compared with the conventional Mata2–Mata1 heterodimer configuration found at haploid-specific genes (Fig. 4). In addition, the spacing between the Mata2 and Mata1 motifs is substantially altered from the conventional scheme (Fig. 4, Supplementary Fig. 2). Model building (based on the existing crystal structures) suggests that Mata1 and Mata2, as they are arranged on the *L. kluyveri* GPA1 regulatory region, could plausibly contact each other (through a short alpha helix on a flexible tether) as is observed in the structure of the conventional heterodimer; however, there must be a severe energetic cost to this altered arrangement because it cannot support repression of GPA1 in the absence of the Mcm1 binding sequence (Fig. 3b) (Li et al. 1995; Tan and Richmond 1998).

The arguments presented above explain, in energetic terms, why all three proteins are needed to repress the *L. kluyveri* GPA1 gene in **a/a** cells. But how might this novel arrangement have evolved? While we cannot provide a definitive answer, there are some important clues buried in the fungal lineage. At the point where *S. cerevisiae* and *W. anomalus* diverge (prior to the divergence of *S. cerevisiae* and *L. kluyveri*), all of the protein–protein interactions needed for the three-part scheme on the *L. kluyveri* GPA1 gene were in place (Britton et al. 2020). Specifically, the favorable contacts between Mata2 and Mcm1 and between Mata2 and Mata1 had evolved before these two branchpoints (Fig. 1b). Thus, the shift between the different modes of regulation could be brought about solely through changes in cis-regulatory sequences. Bioinformatic analysis shows that the conventional form of regulation by the Mata1–Mata2 heterodimer is widely distributed across the ascomycete lineage (Fig. 5, Supplementary Fig. 3). For example, it applies to the haploid-specific genes in *S. cerevisiae*, in *Candida albicans* and (with the exception of GPA1) in *L. kluyveri*. Given its widespread occurrence—especially in species where the Mata2–Mcm1 interaction is absent—the conventional, heterodimer form of regulation almost certainly predates the three-protein mechanism. Accordingly, the three-part form of regulation described here is tightly restricted to a small clade and is most likely a derived form of regulation.

A bioinformatic search, using the three-part *L. kluyveri* sequence described in this paper, failed to identify the *W. anomalus* RME1 control region, indicating that the arrangement of the three proteins differs between these two species. Thus far, our work has documented four different ways in which the haploid-specific genes are regulated across the ascomycete lineage—direct repression by Mata1–Mata2 (as in *S. cerevisiae*), indirect regulation via Rme1 (as in *K. lactis*; (Booth et al. 2010), and two forms of three-part regulation involving Mcm1 [*L. kluyveri* as shown in this work, and *W. anomalus*; (Britton et al. 2020)]. These observations indicate a high degree of flexibility in the way in which the haploid-specific genes are regulated across ascomycete species and raise the possibility that additional mechanisms will be discovered.

This work highlights an important concept in gene expression: the same output (in this case, repression of the haploid genes in **a/a** cells) can be achieved by different mechanistic solutions and—over evolutionary time scales—the mechanism can drift from one solution to another while maintaining the same output. The key to this idea is that gene expression is typically controlled by assemblies of proteins binding cooperatively to control regions on DNA, and the energetics of assembly can be parceled out in different ways, resulting in different types of arrangements on DNA. For example, in the case described here, a deficient binding site for the Mata1–Mata2 heterodimer is compensated by a favorable interaction with a third protein, Mcm1. This idea leads to a cautionary note on interpreting a particular gene expression strategy as somehow perfectly optimized. Instead, as evidenced by comparisons across species, a gene expression scheme is best regarded as a flexible set of possible mechanisms, linked by energetically feasible transitions.

Data availability

Strains and plasmids are available upon request. The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, and files. Complete RNAseq data is available in the Gene Expression Omnibus accession number: GSE221890. Complete ChIP-seq data is available in the Gene Expression Omnibus accession number: GSE221835.

Supplemental material available at GENETICS online.

Acknowledgements

We thank Haley Gause, Candace Britton, Kyle Fowler, Naomi Ziv, Matt Lohse, Trevor Sorrells, Christopher Carlson, Elise Munoz and Eliza Nieweglowska for advice and helpful discussions and Ananda Mendoza for technical assistance. Haley Gause and Jenny Zhang also contributed comments and edits to the manuscript. Sequencing was performed at the UCSF Center for Advanced Technology, supported by UCSF PBBR, RRP IMIA, and NIH 1S10OD028511-01 grants. Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Funding

This work was supported by NIH grant R01 GM037049 (to A.D.J.), and a G. W. Hooper Research Foundation Graduate Student Fellowship (to F.D.F.)

Conflicts of interest

The author(s) declare no conflict of interest.

Author contributions

F.D.F.: Conceptualization, Methodology, Investigation (experimental), Writing—original draft. M.E.G.: Data curation, Visualization, Investigation (in silico), Writing—reviewing and editing. A.D.J.: Supervision, Writing—reviewing and editing.

Literature cited

- Ainavarapu SRK, Brujić J, Huang HH, Wiita AP, Lu H, Li L, Walther KA, Carrion-Vazquez M, Li H, Fernandez JM, et al. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophys J*. 2007;92(1):225–233. doi:10.1529/biophysj.106.091561.
- Andrews. FastQC: a quality control tool for high throughput sequence data; 2010.
- Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Research*. 2015;43(W1):W39–W49. doi:10.1093/nar/gkv416
- Baker CR, Booth LN, Sorrells TR, Johnson AD. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. *Cell*. 2012;151(1):80–95. doi:10.1016/j.cell.2012.08.018.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235–242. doi:10.1093/nar/28.1.235.
- Booth LN, Tuch BB, Johnson AD. Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature*. 2010;468(7326):959–963. doi:10.1038/nature09560.
- Britton CS, Sorrells TR, Johnson AD. Protein-coding changes preceded cis-regulatory gains in a newly evolved transcription circuit. *Science*. 2020;367(6473):96–100. doi:10.1126/science.aax5217.
- Carroll SB. Evo-Devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*. 2008;134(1):25–36. doi:10.1016/j.cell.2008.06.030.
- Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2021;50(D1):D165–D173. doi:10.1093/nar/gkab1113
- Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–i890. doi:10.1093/bioinformatics/bty560.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. Saccharomyces Genome database: the genomics resource of budding yeast. *Nucleic Acids Res*. 2012;40(D1):D700–D705. doi:10.1093/nar/gkr1029.
- Chisanga D, Liao Y, Shi W. Impact of gene annotation choice on the quantification of RNA-Seq data. *BMC Bioinformatics*. 2022;23(1):107. doi:10.1186/s12859-022-04644-8.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCftools. *Gigascience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21. doi:10.1093/bioinformatics/bts635.
- Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, et al. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)*. 2013;4(3):389–398. doi:10.1534/g3.113.008995
- Faber KN, Haima P, Harder W, Veenhuis M, AB G. Highly-efficient electrotransformation of the yeast *Hansenula polymorpha*. *Curr Genet*. 1994;25(4):305–310. doi:10.1007/BF00351482.
- Freese NH, Norris DC, Loraine AE. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics*. 2016;32(14):2089–2095. doi:10.1093/bioinformatics/btw069.
- Galgoczy DJ, Cassidy-Stone A, Llinás M, O'Rourke SM, Herskowitz I, DeRisi JL, Johnson AD. Genomic dissection of the cell-type-specification circuit in *Saccharomyces cerevisiae*. *Proc National Acad Sci U S A*. 2004;101(52):18069–18074. doi:10.1073/pnas.0407611102.
- Gaspar JM. Improved peak-calling with MACS2. bioRxiv 496521; <https://doi.org/10.1101/496521>, 17 December 2018, preprint: not peer reviewed.
- Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci*. 2018;27(1):14–25. doi:10.1002/pro.3235.
- Gojkovic Z, Jahnke K, Schnackerz KD, Piškur J. PYD2 Encodes 5,6-dihydropyrimidine amidohydrolase, which participates in a novel fungal catabolic pathway. *J Mol Biol*. 2000;295(4):1073–1087. doi:10.1006/jmbi.1999.3393.
- Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–1018. doi:10.1093/bioinformatics/btr064.
- Herskowitz I. A regulatory hierarchy for cell specialization in yeast. *Nature*. 1989;342(6251):749–757. doi:10.1038/342749a0.
- Hoffman CS, Winston F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene*. 1987;57(2-3):267–272. doi:10.1016/0378-1119(87)90131-4
- Jarvela AMC, Hinman VF. Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. *Evodevo*. 2015;6(1):3. doi:10.1186/2041-9139-6-3.
- Li T, Stark MR, Johnson AD, Wolberger C. Crystal structure of the MATA1/MATα2 homeodomain heterodimer bound to DNA. *Science*. 1995;270(5234):262–269. doi:10.1126/science.270.5234.262.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.
- Lynch VJ, Wagner GP. Resurrecting the role of transcription factor change in developmental evolution. *Evolution*. 2008;62(9):2131–2154. doi:10.1111/j.1558-5646.2008.00440.x.
- Nocedal I, Mancera E, Johnson AD. Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. *Elife*. 2017;6:e23250. doi:10.7554/eLife.23250.
- Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70–82. doi:10.1002/pro.3943.
- Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. DeepTools2: a next generation

- web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44(W1):W160–W165. doi:10.1093/nar/gkw257.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24–26. doi:10.1038/nbt.1754.
- Shen X-X, Oplente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, et al. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell.* 2018;175(6):1533–1545.e20. doi:10.1016/j.cell.2018.10.023.
- Sorrells TR, Booth LN, Tuch BB, Johnson AD. Intersecting transcription networks constrain gene regulatory evolution. *Nature.* 2015;523(7560):361–365. doi:10.1038/nature14613.
- Sorrells TR, Johnson AD. Making sense of transcription networks. *Cell.* 2015;161(4):714–723. doi:10.1016/j.cell.2015.04.014.
- Génolevures Consortium; Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, Sherman DJ, Weissenbach J, Westhof E, Wincker P, et al. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res.* 2009;19(10):1696–1709. doi:10.1101/gr.091546.109.
- Stern DL, Orgogozo V. The Loci of evolution: how predictable is genetic evolution. *Evolution.* 2008;62(9):2155–2177. doi:10.1111/j.1558-5646.2008.00450.x.
- Strathern J, Hicks J, Herskowitz I. Control of cell type in yeast by the mating type locus the $\alpha 1$ - $\alpha 2$ hypothesis. *J Mol Biol.* 1981;147(3):357–372. doi:10.1016/0022-2836(81)90488-5.
- Tan S, Richmond TJ. Crystal structure of the yeast MAT α 2/MCM1/DNA ternary complex. *Nature.* 1998;391(6668):660–666. doi:10.1038/35563.
- Taylor JW, Berbee ML. Dating divergences in the fungal tree of life: review and new analyses. *Mycologia.* 2006;98(6):838–849. doi:10.1080/15572536.2006.11832614.
- Tsong AE, Miller MG, Raisner RM, Johnson AD. Evolution of a combinatorial transcriptional circuit A case study in yeasts. *Cell.* 2003;115(4):389–399. doi:10.1016/S0092-8674(03)00885-7.
- Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 2012;13(1):59–69. doi:10.1038/nrg3095.
- Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO. Crystal structure of a MAT α 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell.* 1991;67(3):517–528. doi:10.1016/0092-8674(91)90526-5
- Wray GA. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet.* 2007;8(3):206–216. doi:10.1038/nrg2063.
- Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics.* 2012;13(1):134. doi:10.1186/1471-2105-13-134.

Editor: L. Rusche