PNAS RESEARCH ARTICLE | BIOCHEMISTRY OPEN ACCESS

Ancient transcriptional regulators can easily evolve new pair-wise cooperativity

Kyle R. Fowler[a], Fredrick Leon[a], and Alexander D. Johnson[a,b,1]

Edited by H. Mark Johnston, University of Colorado - Anschutz Medical Campus, Aurora, CO; received February 14, 2023; accepted May 26, 2023

Cells regulate gene expression by the specific binding of transcription regulators to cis-regulatory sequences. Pair-wise cooperativity between regulators—whereby two different regulators physically interact and bind DNA in a cooperative manner—is common and permits complex modes of gene regulation. Over evolutionary timescales, the formation of new combinations of regulators represents a major source of phenotypic novelty, facilitating new network structures. How functional, pair-wise cooperative interactions arise between regulators is poorly understood, despite the abundance of examples in extant species. Here, we explore a protein–protein interaction between two ancient transcriptional regulators—the homeodomain protein Matα2 and the MADS box protein Mcm1—that was gained approximately 200 million y ago in a clade of ascomycete yeasts that includes *Saccharomyces cerevisiae*. By combining deep mutational scanning with a functional selection for cooperative gene expression, we tested millions of possible alternative evolutionary solutions to this interaction interface. The artificially evolved, functional solutions are highly degenerate, with diverse amino acid chemistries permitted at all positions but with widespread epistasis limiting success. Nonetheless, approximately ~45% of the random sequences sampled function as well or better in controlling gene expression than the naturally evolved sequence. From these variants (which are unconstrained by historical contingency), we discern structural rules and epistatic constraints governing the emergence of cooperativity between these two transcriptional regulators. This work provides a mechanistic basis for long-standing observations of transcription network plasticity and highlights the importance of epistasis in the evolution of new protein–protein interactions.

gene regulation | transcription | cooperativity | molecular evolution | epistasis

While some transcriptional regulators appear to bind their preferred DNA sites on their own, most bind cooperatively in combination with additional regulators, a property often mediated by weak protein–protein interactions between the regulators (1, 2). The resulting combinatorial control—through which multiple DNA-binding proteins control transcription of a given gene—is a hallmark of gene expression, especially in eukaryotes (3–7). The intrinsic DNA-binding specificity of transcriptional regulators often remains unchanged over long evolutionary times; the extensive (hypothetical) pleiotropy resulting from changes in DNA-binding specificity is thought to significantly constrain such changes (8). However, gains and losses of protein–protein interaction between transcription regulators appear relatively frequently and can occur without extensive pleiotropy (9, 10).

One well-studied gain of a protein–protein interaction is found in a particular clade of fungi where the ancient homeodomain protein Matα2 binds DNA cooperatively with the ancient MADS box protein Mcm1 (Fig. 1*A*) (11). This cooperativity arose approximately 200 million y ago when Matα2 gained the ability to physically interact and cooperatively bind DNA with Mcm1, a change that likely took place only within this clade; this interaction is not found in the species outside the clade (9, 12). The emergence of this cooperative interaction occurred without changes in the DNA-binding specificity of either protein; the newly evolved protein–protein interaction is due to changes only in Matα2 (4, 13, 14). The "recipient" Mcm1 surface involved in the interaction is deeply conserved and did not change when the interaction evolved. The novel Matα2–Mcm1 interaction facilitated the formation of a new transcriptional network to repress the **a**-specific genes, which are central to cell-type specification.

The cooperative interaction between Matα2 and Mcm1 is due to a short (approximately 11 amino acid) region of Matα2 (Fig. 1*B*) (4, 14). A crystal structure of the Matα2–Mcm1 complex bound to DNA shows that this region forms a short beta-strand when bound to Mcm1 (15). Numerous contacts between the two proteins occur over a relatively small (~20 Å) interface, including a cation–pi interaction with a phenylalanine in Matα2 (Fig. 1 *B*, *Inset*). This corresponds to a net cooperative interaction energy of 3 to 4 kcal/

Significance

Changes in gene expression circuits over evolutionary time are a major contributor to the emergence of new phenotypes. Here, we investigate a case where, in the past, two highly conserved transcriptional regulators formed a cooperative interaction such that both proteins are required to bind to DNA and regulate a set of genes. Using deep mutational scanning, we documented many alternative "paths not taken" that give rise to a functional cooperative interaction between these two ancient proteins. We found that functional interactions take a variety of forms and appear to arise with ease. The sheer abundance of alternative solutions to cooperative binding suggests that cooperativity among transcriptional regulators arises frequently over evolutionary time.

Author affiliations: [a]TETRAD Graduate Program, Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158; and [b]Department of Microbiology and Immunology, University of California, San Francisco, CA 94158

Author contributions: K.R.F. and A.D.J. designed research; K.R.F. and F.L. performed research; K.R.F. analyzed data; and K.R.F. and A.D.J. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: ajohnson@cgl.ucsf.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2302445120/-/DCSupplemental.

Published July 3, 2023.

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

mol (11). Prior genetic work (alanine-scanning) has shown that seven contiguous residues in this region of Matα2 (including the above phenylalanine) are essential for efficient combinatorial control of **a**-specific gene repression by Matα2 and Mcm1 (4). Matα2 mediates this repression by recruiting the Tup1-Ssn6 corepressor to DNA through a domain distinct from, and independent of, the Matα2–Mcm1 interaction interface (9).

In this work, we investigate several general questions around the emergence and evolvability of combinatorial gene regulation using the Matα2–Mcm1 complex as a framework. Beginning with constructs missing the pair-wise interaction, we selected—from a highly diverse pool of Matα2 variants—those that could efficiently work in combination with Mcm1 to repress transcription.

We took advantage of the CAN1 gene, expression of which arrests cell growth in the presence of exogenous canavanine (16). We replaced the endogenous CAN1 promoter with a synthetic constitutive construct containing a naturally occurring Matα2–Mcm1 cis-regulatory sequence (Fig. 1*C*). When bound by Matα2–Mcm1, which requires pair-wise cooperativity between the proteins, this sequence element brings about strong transcriptional repression of CAN1 and allows growth in canavanine.

Results

We generated two different Matα2 mutant libraries based on the *Saccharomyces cerevisiae* protein. Both libraries consist of Matα2 from *S. cerevisiae* driven by its endogenous promoter cloned into a low-copy plasmid: Mutations were introduced at key positions that mediate its interaction with Mcm1. The first library introduced all possible individual amino acid substitutions at eleven consecutive residues (G113-M123) known to span the sequence of Matα2 that interacts with Mcm1. Variants in this library possess a single amino acid change relative to the wild-type protein. Seven of these eleven positions (114 to 120) represent

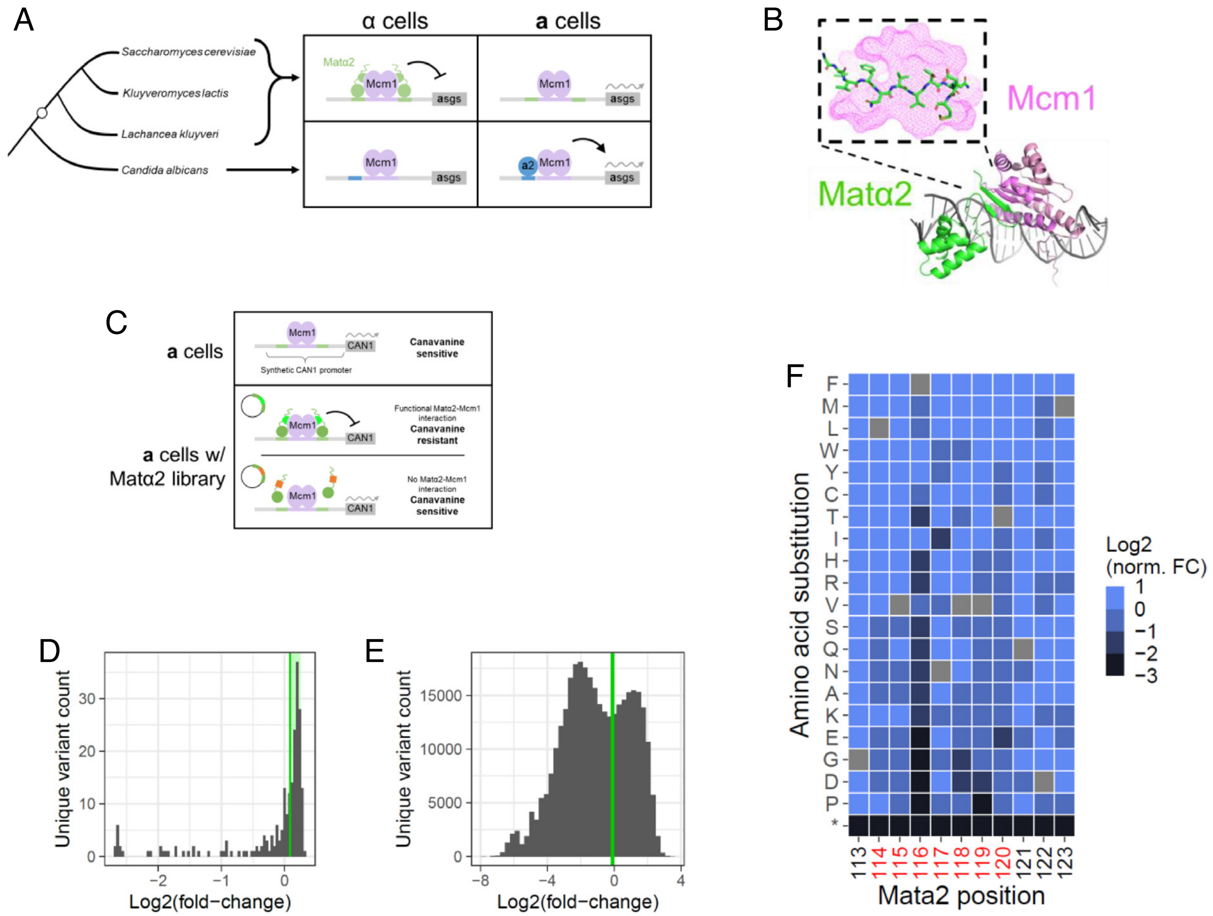

Fig. 1. Evolution of the Matα2–Mcm1 interaction and the strategy for discovering unique functional variants. (*A*) Representative ascomycete yeast species and their mode of **a**-specific gene regulation. Functional Matα2–Mcm1 complexes are found in *S. cerevisiae* and related species due to a short amino acid sequence in Matα2 which interacts with Mcm1 and causes the two proteins to bind DNA cooperatively. Outside the *L. kluyveri* to *S. cerevisiae* clade, the mode of **a**-specific gene regulation is different: An activator made only in **a**-cells induces the genes which are not otherwise expressed (12). (*B*) Structure of the Matα2–Mcm1–DNA ternary complex adapted from Tan and Richmond (1998). The Mcm1 homodimer (magenta and pink) is shown bound by only a single Matα2 protein (green); a second Matα2 (not shown) binds symmetrically to the distal Mcm1 monomer (pink). The eleven amino acids of Matα2 that mediate this interaction (G113-M123) contact Mcm1 directly (*Inset*). (*C*) Selection for new functional Matα2–Mcm1 complexes. CAN1 expression in the presence of canavanine is toxic; repression of the gene by Matα2–Mcm1 allows for normal growth. *S. cerevisiae* **a**-cells (which lack Matα2) transformed with this CAN1 system were susceptible to canavanine even at low concentrations (MIC ~1 μg/mL), while **a**-cells transformed with wild-type Matα2 were able to grow in the presence of high concentrations of canavanine (MIC > 100 μg/mL). (*D*) Fitness distribution of point mutations in the wild-type *S. cerevisiae* Matα2. The histogram shows the fold-changes for all point mutants and controls. Green shaded region is the range of values for wild-type Matα2 and variants with synonymous mutations. Wild-type *S. cerevisiae* Matα2 is indicated by the dark green vertical line. (*E*) Fitness distribution of randomized Matα2 variants. Each protein variant contains seven random residues at positions 114 to 120. The green line indicates the fold-change of the wild-type Matα2 from *S. cerevisiae*. Those to the right of the line show greater enrichment than that of the wild-type sequence. (*F*) Heat map showing fold-change values for point mutants of the wild-type *S. cerevisiae* Matα2. The effects of the indicated amino acid substitution (*y* axis) at the indicated positions (*x* axis) are colored by fold-change after selection and normalized to the wild-type (set at zero). The seven core residues are indicated in red. Premature stop codons are indicated by *. Gray boxes denote the wild-type amino acids. Note that two mutants—N117I & N117V—exhibited a similar growth defect in the absence of canavanine selection, suggesting that these specific mutational effects do not involve Mcm1.

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

the essential "core," which interacts directly with Mcm1 as observed in a crystal structure of the Matα2–Mcm1 complex bound to DNA (15). A wild-type Matα2 construct (Sc-Matα2) was also included as a control.

In the second library, we randomized the seven core residues using an NNK oligonucleotide strategy and obtained ~1.2 million constructs with unique amino acid combinations. Variants in this library differ from the wild-type protein, on average, at all seven amino acids. With this library, we sought to reveal the ease with which functional protein–protein interactions could evolve. The total number of possible amino acid combinations in this library is immense (20^7 or ~ 1.3 billion possibilities) and technically infeasible to sample completely. However, we reasoned that even a sparse sampling of this sequence space could reveal important functional trends.

To screen for functional variants of Matα2 that could interact with Mcm1, we transformed the Matα2 mutant libraries into *S. cerevisiae* **a**-cells where the only expressed Matα2 protein is from the plasmid libraries. As a control, we spiked in cells carrying the wild-type Sc-Matα2 plasmid at a concentration of 0.1%. Cells with these Matα2 constructs were then grown for 24 h in media either lacking canavanine (representing the pool of unselected variants) or at 250 µg/mL to enrich for functional variants that could repress transcription of CAN1. Following growth, which corresponds to ~5 generations, the Matα2 plasmid was purified from the final populations and sequenced deeply. The frequency of any given construct among the reads from each pool correlates with its abundance in that population of cells. We could then calculate a fold-change (FC) for each construct after selection relative to either the starting population or the population grown without canavanine selection, allowing us to control for growth effects that are independent of canavanine. For example, an FC of < 1 means that the sequence was selected against in canavanine. This fold-change allows us to estimate relative fitness for each sequence examined.

In both libraries, Matα2 mutants with premature stop codons diminished in frequency in the population and therefore exhibited low FCs. Conversely, the frequency of the wild-type Sc-Matα2 gene—as well as those bearing synonymous mutations—remained stable after selection (Fig. 1 *D* and *E*). The fully functional Sc-Matα2 had an FC around 1.0 and was surpassed by numerous other constructs. This indicates that our selection regime successfully enriched for functional Matα2–Mcm1 interactions.

Using the first library, we assessed the consequences of single amino acid mutations on Matα2–Mcm1 function. Many amino acid substitutions were well tolerated and remained at high frequency after selection (Fig. 1*D*), a pattern not necessarily predicted from the alanine-scanning experiments (4). Alanine substitutions in our library had modestly reduced FCs, especially at the "core" seven positions, a result entirely consistent with previous work.

For example, the phenylalanine at position 116 is especially critical, as many substitution mutations were highly detrimental (Fig. 1*F*). However, replacement with another large aromatic (e.g., F116W) or aliphatic residue (F116I) did not diminish its function, suggesting that a variety of bulky or hydrophobic side chains at this position would suffice for the Matα2–Mcm1 interaction. Aromatic amino acids were also tolerated at other positions throughout the region: mutation to phenylalanine, tryptophan, or tyrosine at most positions had little or no effect. Together, these results suggest that the Matα2–Mcm1 interaction depends broadly on hydropathy, with many different individual sequences sufficing for function.

This view was supported and greatly extended by the results of our second library where the interfacial positions of Matα2 were randomized, generating a complex library of amino acid sequences. Following the selection scheme described above, these Matα2 variants exhibited a broad, largely symmetric fitness distribution with a median FC around one (Fig. 1*E*). The spiked-in wild-type construct also exhibited an FC close to one (0.93) but was surpassed by many variants, indicating a large dynamic range of "successful" variants, many of which outperform the natural sequence under the selection scheme imposed. Using the naturally occurring Sc-Matα2 as a basis for comparison, we initially estimate that ~35% of variants in our library (> 100,000 unique proteins) are at least as functional as the extant protein in our transcriptional repression assay (Fig. 1*E*). When we exclude variants with low starting frequencies (that is, those with lower statistical significance), the percentage of fit sequences was even higher approaching ~45% (*SI Appendix*, Fig. S1*A*). Thus, many combinations of random amino acids result in a functional Matα2–Mcm1 complex, suggesting that a surprising fraction of this sequence space is functional. We refer to these as "fit" Matα2 variants. The high frequency of fit variants was reproducible between replicates and robust to read depth (*SI Appendix*, Fig. 1 *B* and *C*).

What is the molecular basis for this large number of functional, fit alternatives? Do their solutions exhibit any patterns? Normalizing amino acid frequencies in fit sequences to their abundances in the unselected library revealed how residues increased or decreased in frequency after selection, and therefore which residues promote function. Phenylalanine showed the strongest enrichment at every position, followed by most aliphatic amino acids and tryptophan (Fig. 2*A* and *SI Appendix*, Fig. S1 *D* and *E*). Disfavored amino acids were primarily charged. Most striking, however, is the lack of position specificity: for example, phenylalanines are broadly beneficial at each of the seven randomized positions. Nonetheless, as discussed below, the quantitative effects are highly dependent on the amino acids in the remaining positions.

When we considered combinations of residues, rather than individual amino acids, we detected some additional patterns among fit sequences; for example, although a single phenylalanine promotes fitness at all positions (Fig. 2*A*), two phenylalanines occurred less frequently together than expected by chance (Fig. 2*B*). Nonadditive interactions, typified by this observation, are indicative of intramolecular epistasis, and we next examined epistasis more systematically. For example, among variants with a phenylalanine at the seventh position (7F), the pattern of favored/disfavored amino acids resembles that in the general population (Fig. 2*C*). But among the fit 7F sequences, all previously favorable amino acids (e.g., aromatic residues) occurred less frequently, while disfavored amino acids (e.g., charged residues) were more abundant (Fig. 2*D*). Even prolines, which were generally selected against and known to be disruptive to protein secondary structure, were enriched among 7F variants. This pattern was highly idiosyncratic, however, with drastically different amino acid biases exhibited when the phenylalanine was fixed at different positions, or when other amino acids were fixed (*SI Appendix*, Fig. S2 *A* and *B*). We further probed the extent and nature of epistasis by quantifying pair-wise interactions between all amino acid states across the seven positions. To do so, we first determined the frequency of each amino acid at each position among the fit variants (Fig. 3*A*). We then calculated the expected co-occurrence of each amino acid state pair assuming independence (i.e., no epistasis). Deviations from this expectation indicate a genetic interaction between residues (Fig. 3*B*).

We found numerous instances of both positive and negative epistatic interactions (Fig. 3*B*). For example, the heat map illustrating

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

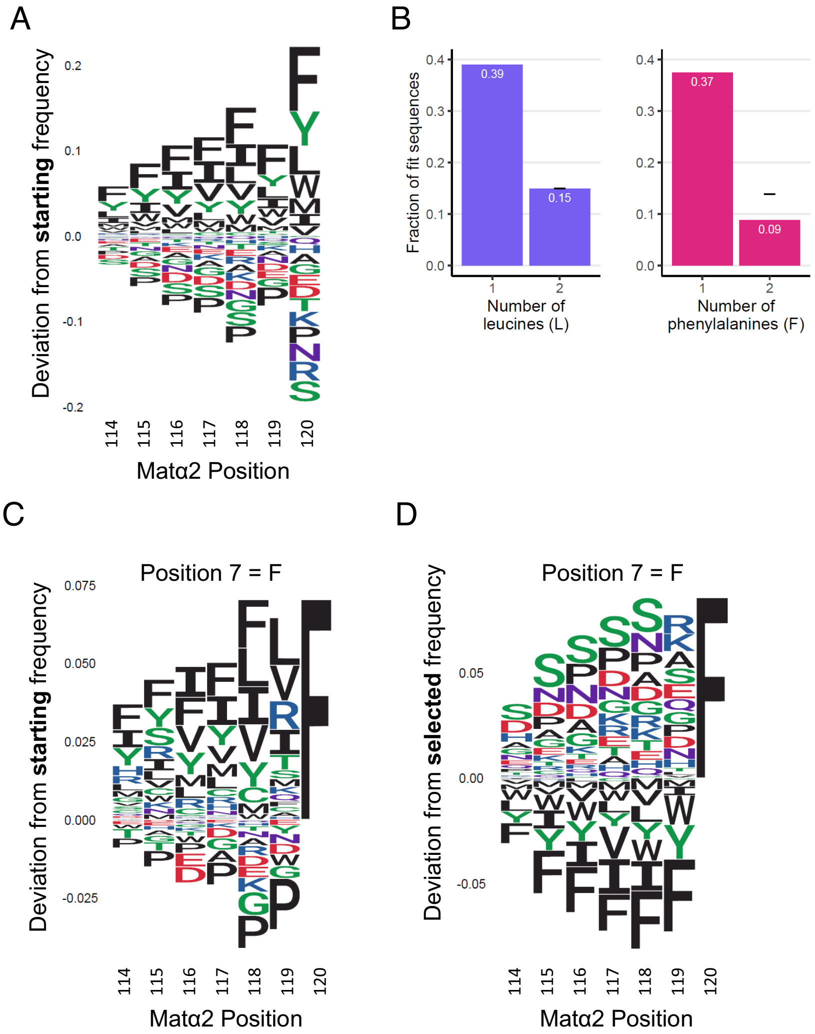

Fig. 2. Functional Matα2 proteins with highly degenerate interaction interfaces. (*A*) Normalized logo plot showing changes in amino acid abundance at each position after selection. Positive values indicate amino acids that increased in frequency, while frequency decreases are negative values. The size of each letter corresponds with the magnitude of the change and is ordered from the largest (*A*, *Top*) to the smallest (*A*, *Bottom*) change. Phenylalanine is universally beneficial at all positions: F114 was slightly enriched, while F120 is strongly favored. Selection against disfavored residues was stronger at this latter position than that at other positions. (*B*) The occurrence of multiple phenylalanines is underrepresented among functional variants based on the frequency of single occurrences. The fraction of fit variants with a single leucine or phenylalanine (left columns) was used to calculate the expected frequency of double leucine or phenylalanine variants (black horizontal line). Note that while the number of functional variants containing two leucines is approximately the square of the single leucine frequency, double phenylalanine variants are less abundant than expected if they were independent. (*C*) Favorable amino acid compositions are heavily context dependent. Normalized logo plot as in *A* generated from the subset of variants with a phenylalanine at position 120 (F120). Amino acid frequency changes are relative to their preselection frequencies. (*D*) Influence of F120 on adjacent amino acids. Logo plot of F120 sequences normalized to postselection amino acid frequencies showing the additional effect of F120 relative to other selected sequences. The influence of F120 on amino acid composition contrasts strongly with the overall pattern in *A*.

the relationship between the first and seventh positions reveals that many amino acid pairs exhibit some degree of epistasis (Fig. 3*C* and *SI Appendix*, Fig. S3*A*). This pattern was consistent between replicate selections and was not observed in the absence of selection (Fig. 3*B* and *SI Appendix*, Fig. S3 *B* and *C*). In sum, 29% of all pairs of amino acid states along the interface exhibited a significant epistatic interaction, with positive and negative interactions equally represented.

Finally, we considered whether the fitness landscape of functional variants was rugged or smooth. Based on the effects of single amino acid substitutions away from fit variants, we conclude that most fitness peaks were relatively broad, with many single substitutions showing little or no decreased fitness (Fig. 4*A* and *SI Appendix*, Fig. S4). The magnitude of the fold-change difference between our matched mutational pairs describes peak "steepness," while the landscapes overall "smoothness" can be estimated from the fraction of all mutation pairs where function is compromised by the mutation. Overall, only ~20% of matched mutation pairs exhibit this pattern, while specific substitutions (e.g., position 120 to phenylalanine) alter fitness more frequently (Fig. 4*B*).

Our analysis has relied thus far on the ability of Matα2 variants to repress transcription of an artificial CAN1 reporter construct, and here we consider the natural function of Matα2. We tested several variants spanning a range of FC values by replacing the endogenous Matα2 locus with the variant and monitoring the normal role of Matα2, namely, in promoting the ability to mate, which requires repression of multiple **a**-specific genes by Matα2, not just a single gene as in our selection. Using a quantitative mating assay, we measured the mating efficiencies of α-cells bearing either wild-type or mutant Matα2. Cells with wild-type Matα2 were highly proficient at mating (*SI Appendix*, Table S1). In contrast, and as a control experiment, replacing the Mcm1 interaction region of Matα2 with that of *C. albicans*, which diverged prior to the emergence of the interaction, strongly reduced mating efficiency. Replacement with variants from our library resulted in a range of mating efficiencies.

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

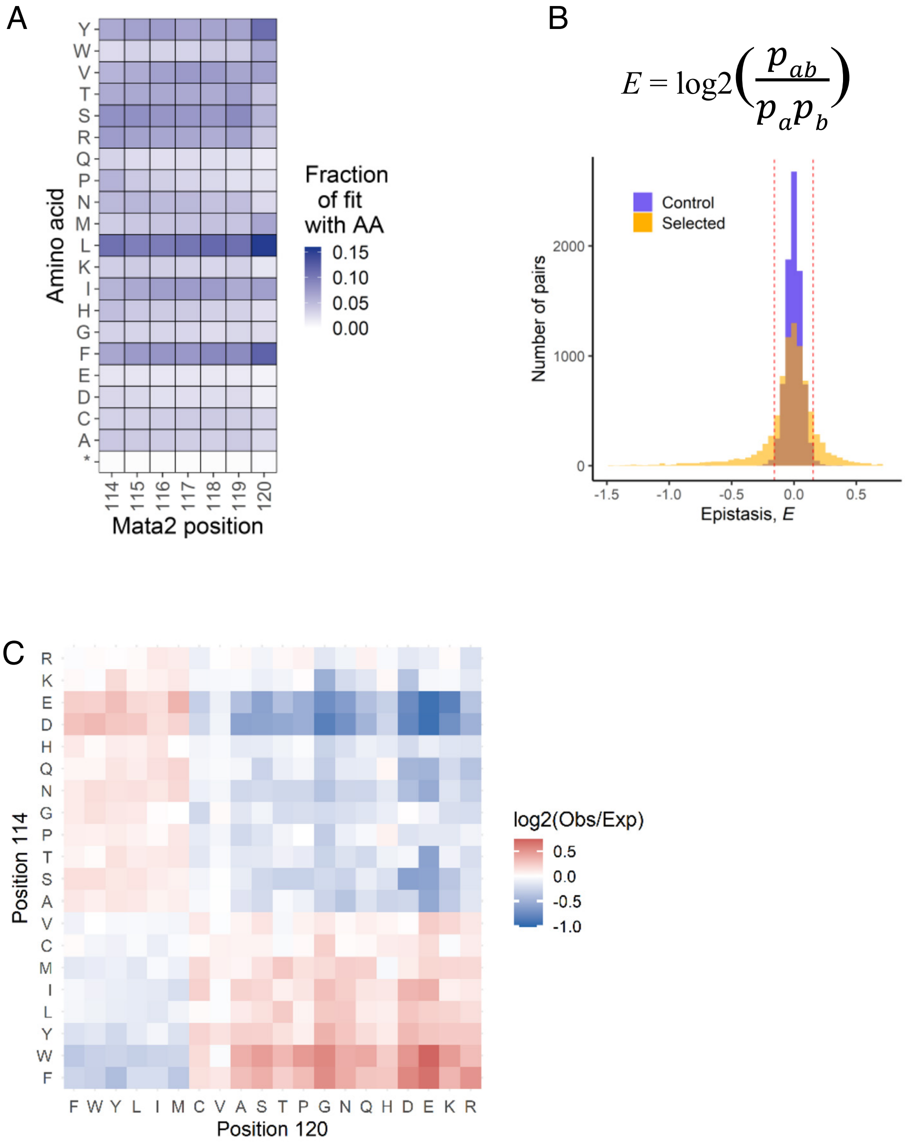

Fig. 3. Functional Matα2–Mcm1 interactions exhibit extensive intradomain epistasis. (*A*) The frequency of amino acids across all positions among functional variants. Stop codons are indicated by *. (*B*) Histogram of the effects of epistasis on the frequency of amino acid pairs across all positions relative to expected frequencies given independence. Values were calculated using unselected control sequences (purple) to define a 95% CI (red dashed lines). (*C*) Heat map showing the frequency of amino acid pairs at positions 114 & 120 relative to their frequency, assuming complete independence as in *B*.

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

Cells bearing Matα2 variants with low FC scores mated as poorly as the *C. albicans*-like protein. Conversely, the mating efficiencies of the highest scoring variant (FC = 5.2, core sequence PCLRFVF) mated as well as wild-type Matα2. However, a variant with an intermediate FC of 1.6, which indicates proficient growth in the bulk canavanine competition, was not able to restore mating. This discrepancy likely reflects the different requirements of these assays: Repression of the **a**-specific genes for mating involves the binding of Matα2–Mcm1 to a range of cis-regulatory sequences at multiple genes, while growth in canavanine involves binding a single DNA sequence at CAN1. However, we emphasize that a random Matα2 variant, chosen for repressing our CAN1 reporter, also functions as well as the wild-type Matα2 in its natural setting to repress the **a**-specific genes and allow mating (*SI Appendix*, Table S1).

Discussion

In this work, we investigated how a pair of deeply conserved ancient transcriptional regulators can acquire a protein–protein interaction that results in their ability to cooperatively and efficiently regulate gene expression. The framework for the study is based around a known protein–protein interaction between an ancient homeodomain protein (Matα2) and an ancient MADS domain protein (Mcm1) that evolved relatively recently in the *S. cerevisiae* clade. By investigating a large number of the "paths not taken," we uncovered the constraints—or seemingly lack thereof—governing the de novo emergence of a functional interaction that allows these two ancient proteins to work in combination to regulate gene expression. Our study shows that many different solutions in the Matα2 protein are capable of mediating a functional, cooperative interaction with Mcm1. We further probed the chemically diverse set of novel Mcm1-interacting interfaces to reveal rules and constraints, including widespread intramolecular epistasis, underlying this interaction.

Our principal conclusions are as follows:

1. Many alternatives to the "naturally evolved" sequence function as well or better, suggesting that functional cooperative interactions between transcriptional regulators can arise with relative ease. Remarkably, nearly half of the randomly generated,

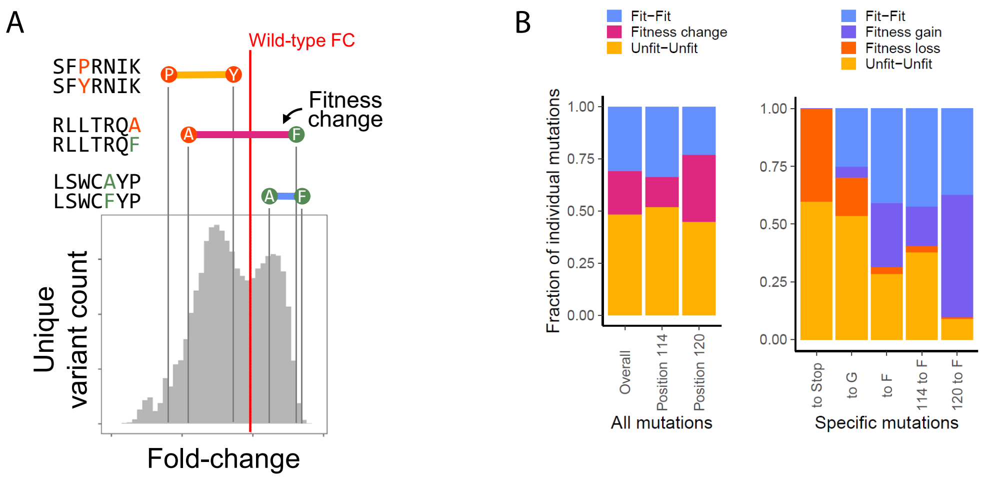

Fig. 4. Fitness landscape topology. (*A*) Effects of single amino acid mutations based on matched pairs of variants that differ by a single residue. Because this sequence space is so vast, we relied on our library's sparse but broadly sampled set of ~11,000 variant pairs that differ by a single amino acid. Each pair represents a single, randomly sampled mutational step from across the entire fitness landscape. (*B*) Individual mutational steps reveal a smooth landscape. Variants that differ by a single residue were grouped based on the position of the difference (*B*, *Left*) or by specific mutations (*B*, *Right*, e.g., mutating any position to glycine or position 120 to phenylalanine). Changes in fitness were defined by the FC of wild-type Matα2. Matched variants with phenylalanine at position 120 were functional 90% of the time (by either retaining or gaining functionality) when compared with a matched sequence with a different amino acid only at this position.

de novo derivatives we assayed worked as well or better than the natural sequence.

2. Successful interactions between the two transcriptional regulators studied here are a distributed property of the interface. For example, a single phenylalanine at the interaction surface is broadly beneficial at every position; however, once a phenylalanine is fixed in position, severe epistasis constrains all other positions.
3. The consequences of single point mutations away from a naturally occurring sequence (represented by our first deep scanning library) did not predict the high degree of idiosyncratic solutions (found in our second, de novo library) that bear little resemblance to the natural sequence.

How do the results presented here compare with other studies? When mutations are considered in combination, epistasis has frequently emerged from studies of protein evolution (17–21). Notably, the evolution of other pair-wise protein–protein interactions has revealed instances of intramolecular epistasis. For example, particular residues in the interface between the *E. coli* PhoQ protein kinase and its substrate PhoP exhibit epistatic interactions (18). Likewise, the evolution of bacterial toxin–antitoxin proteins appears to frequently involve intermediate protein states that are promiscuous, allowing antitoxins to simultaneously recognize and inactivate an ancestral cognate toxin as well as a newly evolved toxin target (17). Such promiscuous intermediates facilitate the evolution of new toxin–antitoxin specificities by avoiding nonfunctional (or less-functional) intermediates. The epistasis we observe between residues of Matα2 likewise appears to facilitate the emergence of a new function, namely interaction with Mcm1. The extreme degree of epistasis observed here, however, is much greater than previously observed. We found that every position of the interface, for multiple amino acid states, exhibits epistasis. Other instances of intramolecular epistasis have involved one or a few positions, and typically between specific amino acids (17, 18, 20, 22). Selection for high binding specificity, as in two-component and toxin–antitoxin systems in order to avoid "crosstalk" with numerous paralogs, may also limit intramolecular epistasis.

Although for our study Matα2 and Mcm1 are positioned precisely on their adjacent DNA sites, the wide range of functional Matα2–Mcm1 interactions identified in our study are reminiscent of the "fuzzy" interactions between the activation domains (ADs) of some transcription factors that bind subunits of the Mediator complex (23–26). Akin to the Matα2–Mcm1 interaction, this "fuzzy" binding involves degenerate interfaces that are enriched for hydrophobic residues and favor aromatic residues (27–29). Besides bulky hydrophobic residues, ADs are also rich in negatively charged residues; however, the opposite preference is observed among Matα2 variants able to function with Mcm1. These trends suggest that AD- and Matα2-mediated interactions have certain similarities but also important differences. More specifically, the origin of the "fuzziness" is likely to be different: ADs have the possibility of interacting with many different sites on Mediator—with few spatial constraints. In contrast, Matα2 and Mcm1 must interact with precise geometry to maximally satisfy the protein–DNA interaction, which also necessitates DNA bending (15, 27, 28). It is also likely that, due to the thermodynamic contributions of DNA binding, a functional Matα2–Mcm1 interaction could be weaker than those between ADs and Mediator, contributing to the different range of possible solutions.

Finally, we bring up an apparent paradox raised by our results: If so, many solutions exist for a functional interaction between Matα2 and Mcm1, why is the naturally occurring solution preserved across a clade of closely related species? We consider three possibilities. 1) There is some other function for this region of Matα2 that constrains its sequence. This possibility seems unlikely given the detailed biochemical, genetic, and structural studies of Matα2 over the past three decades, which has revealed that the only role of this stretch of amino acids considered here is interacting with Mcm1 (4, 11, 13). The fact that one of our randomly chosen de novo variants (which lacked any resemblance to the naturally occurring sequence) functioned and properly recapitulated the normal role of Matα2–Mcm1 in promoting mating argues against an "unknown" function that would have a noticeable consequence on the cell. 2) The Mcm1-interacting sequence of the wild-type Matα2 is constrained to prevent its promiscuous interactions with other transcriptional regulators. Over the short term, we know that the de novo derived functional Matα2s do not cause a noticeable fitness defect; that is, their representation in the control experiment (in the absence of canavanine) did not increase or decrease. However, it is possible that over long

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

evolutionary times, small defects caused by promiscuous Matα2s could constrain its sequence. 3) The naturally occurring sequence arose historically and even small changes which weaken or strengthen the interaction are selected against over long evolutionary times. Irrespective of the answer to this question, our results show that—at least in the short term—many different solutions are possible to productively join two transcription regulators.

In conclusion, the purpose of this study was to explore the range of possible solutions that can make a functional protein–protein interaction between two ancient transcription regulators, resulting in their working cooperatively to control gene expression. Our finding that approximately 45% of random amino acid sequences can recapitulate a functional interaction between two ancient proteins suggests that new interactions between existing transcriptional regulators are sampled continuously over evolutionary time, some of which are ultimately retained by selection. We suggest that the weak, degenerate nature of the pair-wise cooperative interaction observed here is broadly applicable and could facilitate the relatively rapid rewiring of transcription networks and the consequent arrival of new phenotypes.

Materials and Methods

***S. cerevisiae* Strain Construction.** All *S. cerevisiae* strains were in the S288C background and grown on yeast extract peptone dextrose (YEPD) media at 30 °C unless otherwise indicated. Transformations were conducted using the standard lithium acetate/polyethylene glycol method (30). In the S288C **a**-cell, the CAN1 gene was engineered to be repressed by Matα2–Mcm1 by inserting immediately upstream of the CAN1 ORF a PCR product amplified from pKF145 using oKF437 and oKF438. This PCR product contains part of the CYC1 promoter with a Matα2–Mcm1 cis-regulatory sequence from STE2 inserted upstream of its transcriptional start site. The resulting strain yKF230 constitutively expresses CAN1 in the absence of Matα2 but strongly repressed by Matα2–Mcm1. Deletion of the silent MATα locus (HML) in yKF230 used an NatR marker amplified from pFA6a-natMX with homology to HML and resulted in yKF231. The initial Matα2 selection screen was done in yKF230 (in which HML is intact), with all subsequent work being done in yKF231 (HMLΔ). All genetic manipulations were confirmed by PCR and DNA sequencing.

Matα2 Expression Plasmid Construction. Matα2 and its endogenous promoter were synthesized as gBlocks (IDT) and ligated together before being inserted into the AscI site of pRNDM (Addgene), a compact CEN/ARS plasmid, to generate pKF146. This Matα2 expression plasmid was subsequently digested with NdeI and AgeI and oKF457 was inserted to generate a mutant Matα2 with a unique EcoRV site in place of the Mcm1 interaction region (pKF154). This served as an efficient "landing pad" for the subsequent insertion of various DNA sequences, which eliminates the EcoRV site in the process of regenerating either the wild-type Matα2 DNA sequence or variants containing mutations. A separate silent mutation was also introduced nearby at I124 (codon ATA to ATC) to differentiate the Matα2 construct from any chromosomal gene sequence. Regeneration of full-length Matα2 was accomplished using the NEBuilder HiFi DNA Assembly master mix (New England Biolabs) following EcoRV digestion of pKF154. The wild-type *S. cerevisiae* Matα2 protein was assembled using oKF458 and oKF459. To test whether the homologous region of Matα2 from *C. albicans* was capable of interacting with Mcm1, a chimeric protein was constructed using oKF460 and oKF461 which substitutes the *C. albicans* amino acid sequence SPFSNSADT in place of the *S. cerevisiae* sequence GLVFNVVTQDM.

Design of Matα2 Libraries. The assembly of Matα2 mutant libraries was conducted as above using NEBuilder HiFi DNA Assembly master mix (New England Biolabs) except that degenerate oligonucleotide pools synthesized by Integrated DNA Technologies (IDT) were used. Two different mutant libraries were generated. The first consisted of single amino acid changes at each position in the *S. cerevisiae* Mcm1 interaction region. This was accomplished by annealing pairs of oligonucleotides (oKF521-542) in which each codon in the 11 amino acid Mcm1 interaction region has been separately replaced by an NNK codon (where N indicates an A, C, G, or T, and K indicates a G or T). This resulted in 11 separate plasmid pools (pKF159-169) each with 32 possible DNA sequences at each NNK codon (4 × 4 × 2). In the second library, a single pair of annealed oligonucleotides (oKF462 and oKF463) containing seven consecutive NNK codons was used to randomize positions L114-T120, generating library pKF157. This second library consists of many distinct Matα2 variants, each with all seven core residues randomized, and the total possible number of combinations (327) exceeding 34 × 109.

Following assembly, these Matα2 mutant constructs were electroporated into 5-alpha electrocompetent *Escherichia coli* cells (New England Biolabs) according to the manufacturer's instructions. The first library of codon point mutations were transformed separately for each codon and the cells then pooled. To maximize the number of transformants from the second (7 × NNK) randomization library, ten electroporations were done in parallel and later pooled. Immediately after electroporation, prewarmed SOC media were added to the cells, and they were recovered for 1 h at 37 °C. Following recovery, we determined the transformation efficiencies and library complexities by diluting an aliquot of cells and plating on LB media supplemented with 30 μg/mL kanamycin. The single-codon, point mutation library reached saturation, with every DNA mutation being represented several times over in the transformed stock. We estimate the complexity of the second 7 × NNK library to be ~2.92 million unique transformants.

To select *E. coli* transformants in bulk, each library was used to inoculate 500 mL LB media supplemented with 50 μg/mL kanamycin and grown overnight shaking at 37 °C. The next morning, plasmid was purified from each culture using ten QIAprep spin miniprep columns (QIAGEN). The saturated overnight *E. coli* cultures were also used to make 1 mL and 10 mL library freezer stocks stored at −80 °C for later use.

The purified plasmid libraries were used to transform yKF230 or yKF231 to G418R according to a high-efficiency yeast transformation protocol (31). Prior to transformation, a plasmid bearing the wild-type *S. cerevisiae* Matα2 was spiked into both libraries at a frequency of 1/1,000. For the more complex 7 × NNK library, six transformations were done in parallel and then pooled. Transformants were selected in bulk by adding each transformation to 500 mL YEPD media supplemented with 200 μg/mL G418 sulfate and growing overnight shaking at 30 °C. An aliquot of transformed cells were also diluted and spread on YEPD plates containing 200 μg/mL G418 sulfate to determine transformation efficiencies. Transformation with the point mutation library resulted in ~1.1 million transformants; the pooled 7 × NNK library had ~1.88 million transformants. Following overnight growth, 5 mL of saturated yeast culture was combined with 5 mL 50% glycerol to make freezer stocks, which were subsequently stored at −80 °C for later use.

Canavanine Selection Assay and Sequencing. Frozen yeast library aliquots were thawed on ice and then added to 500 mL YEPD. Library cultures were grown overnight shaking at 30 °C to allow for the equal expansion of all Matα2 variants. The following morning, cells were collected for sequencing by pelleting 45 mL of saturated culture, washing once in PBS, and then freezing the cell pellets at −80 °C for later plasmid purifications. This is the "Pre" library and provides the starting frequencies of each variant. Selection for Matα2 variants capable of functionally interacting with Mcm1 was carried out by diluting 5 mL of saturated overnight culture into 500 mL synthetic medium lacking arginine and supplemented with either 25 or 250 μg/mL L-canavanine ("low" and "high" treatment conditions, respectively). To control for growth differences due to the plasmid or protein expression, the same media lacking L-canavanine were used to start a third culture ("zero" treatment) in which there is no selection on Matα2–Mcm1 function. After 24 h of growth in either zero, low, or high canavanine, cells were pelleted in 45 mL aliquots and the pellets frozen at −80 °C. Note the low and high canavanine selections resulted in very similar results and thus we focused our analysis on the high selection condition. All figures use the high canavanine selection data and any conclusions were corroborated using the low canavanine data.

The Matα2 plasmid library was recovered by boiling the cell pellets for 5 min and then bead-beating in the presence of phenol and chloroform. DNA was ethanol precipitated and purified using a QIAprep miniprep kit (QIAGEN). To prepare sequencing libraries, PCR was used to amplify the variable region of Matα2 from the plasmid pool using primers containing Illumina adapter sequences and sample specific barcodes. For each library, ten 50 μL PCR reactions were pooled after nine PCR cycles. Each 500 μL PCR pool was cleaned and concentrated using a MinElute PCR Purification Kit (QIAGEN) and quantified using either a Bioanalyzer or TapeStation (Agilent). Massively parallel sequencing was carried out using single-end 50 base pair reads on an Illumina HiSeq 4000 sequencing system.

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

Downloaded from https://www.pnas.org by UC San Francisco on May 8, 2024 from IP address 169.230.7.236.

Data Processing and Analysis. Processing of sequence data was carried out in the programming language R using Bioconductor and custom scripts. Sequence reads were first trimmed based on their Phred quality scores to remove bases distal to the point a read's quality drops below 20 (corresponding to 1% error). Trimmed sequences were then aligned to the Matα2 reference sequence, and reads were eliminated if they did not fully span the mutated region. Furthermore, reads were removed if they contained indels or any mutation outside the mutated region. Sequences passing these quality filters were tabulated and the counts normalized to the sequencing depth (i.e., reads per million). The normalized counts for each unique Matα2 sequence were then compared across all conditions. Fold-changes were calculated by taking the ratio of the normalized counts after selection over the frequency prior to selection (or without selection using the zero canavanine data). Matα2 variants were eliminated from most analyses if the variant was sequenced less than 10 times either before or after selection.

Quantitative Mating Assays. Quantitative mating assays were performed according to previously described methods for *S. cerevisiae* (32). Strains to test were created by replacing the endogenous Matα2 locus with specific variants identified from our library. The chosen variants span the full range of fold-change values and were highly reproducible between replicate experiments. Specific mutations were first cloned into pKF154, the Matα2 expression plasmid used above, as previously described to generate the mutant library. Plasmids with these new variants were digested with NdeI and NruI to liberate a DNA fragment containing Matα2 and its promoter and a KanMX resistance marker. PCR primers were used to add sequence homology for the mating-type locus to the end of the KanMX gene. These DNA fragments were then used to transform yKF249, an α-cell from the W303 background with Matα2 replaced by the URA3 gene. As controls, we also introduced the wild-type *S. cerevisiae* Matα2 gene and a variant with the *C. albicans* region which is unable to interact with Mcm1. All genotypes were confirmed by PCR and sequencing.

The α-cells created above bearing Matα2 variants were Trp- G418R and were mated to Trp+ **a**-cells. For each mating, the strains were grown to mid-log phase and their OD600 was measured. Cells were then combined with an a:α ratio of 10:1 and concentrated onto 0.8 μm nitrocellulose filters using a Millipore 1,225 Vacuum Sampling Manifold. The filters were then placed on YEPD agar plates and incubated for either 4 or 16 h at 30 °C to allow mating. The filters were then vortexed in 5 mL water to resuspend cells for plating. Dilutions were first plated on YEPD plates containing 200 μg/mL G418 sulfate and grown for 2 d at 30 °C to select for conjugants and the limiting parental strain. These G418R colonies were counted and then replicated to SD-Trp to select for conjugants only. The Trp+ G418R colonies were counted and mating efficiencies calculated as follows: Mating efficiency = (number of Trp + G418R colonies) / (total number of G418R colonies).

Growth Competition Assays. Two Matα2 variants from our library appeared to grow slowly in the absence of canavanine, suggesting that some Matα2 mutations may be detrimental to growth. Both these variants involve a single amino acid change to the wild-type *S. cerevisiae* sequence and are at the same position (N117I and N117V). To test whether these specific mutations impact growth, we introduced them into the endogenous Matα2 locus as above for quantitative mating assays. We then carried out growth competitions in an attempt to measure even subtle growth differences. Each mutant was competed against an isogenic parental strain bearing wild-type Matα2 and constitutively expressing mCherry. For each competition, the mutant strain (N117I or N117V) and mCherry competitor were grown separately to saturation in liquid YEPD. Their OD600 was then measured and combined at a 1:1 ratio in a final volume of 1 mL. This mixture was then diluted to OD600 = 0.1 in synthetic (SD) media lacking arginine, which was the media condition in which the slow growth was first observed, and grown overnight at 30 °C. The cultures were back diluted to OD600 = 0.1 the next morning and grown again overnight at 30 °C. This repeated growth and dilution process continued for 5 d. With each daily passaging, cells were also removed and counted on a BD FACSCelesta flow cytometer. The relative growth rate of the nonfluorescent mutant strain and the mCherry wild-type strain was then determined.

Data, Materials, and Software Availability. DNA sequences (FASTQ files) data have been deposited in Gene Expression Omnibus (GSE233191 (33)).

ACKNOWLEDGMENTS. We thank A. Phillips and N. Ziv for valuable comments on the manuscript and N. Ziv, M. Lohse, J. Fraser, and H. Madhani for critical experimental advice. We also thank C. Britton, L. Noiman, and F. Del Frate for helpful discussions during the conception of this project. This work was supported by NIH grant R01 GM037049 (to A.D.J.) and an NSF Graduate Fellowship (to K.R.F.). Portions of this paper were developed from the thesis of K.R.F.

1. T. R. Sorrells, A. D. Johnson, Making sense of transcription networks. *Cell* **161**, 714–723 (2015).
2. W. Wang *et al.*, Inference of combinatorial regulation in yeast transcriptional networks: A case study of sporulation. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1998–2003 (2005).
3. A. Jolma *et al.*, DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
4. J. Mead, H. Zhong, T. B. Acton, A. K. Vershon, The yeast alpha2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. *Mol. Cell. Biol.* **16**, 2135–2143 (1996).
5. K. Monahan *et al.*, Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *ELife* **6**, e28620 (2017).
6. Y. Pilpel, P. Sudarsanam, G. M. Church, Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29**, 153–159 (2001).
7. T. Ravasi *et al.*, An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).
8. G. A. Wray, The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007).
9. C. S. Britton, T. R. Sorrells, A. D. Johnson, Protein-coding changes preceded cis-regulatory gains in a newly evolved transcription circuit. *Science* **367**, 96–100 (2020).
10. T. R. Sorrells *et al.*, Intrinsic cooperativity potentiates parallel cis-regulatory evolution. *ELife* **7**, e37563 (2018).
11. A. K. Vershon, A. D. Johnson, A short, disordered protein region mediates interactions between the homeodomain of the yeast α2 protein and the MCM1 protein. *Cell* **72**, 105–112 (1993).
12. C. R. Baker, L. N. Booth, T. R. Sorrells, A. D. Johnson, Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. *Cell* **151**, 80–95 (2012).
13. J. Mead *et al.*, Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. *Mol. Cell. Biol.* **22**, 4607–4621 (2002).
14. A. E. Tsong, B. B. Tuch, H. Li, A. D. Johnson, Evolution of alternative transcriptional circuits with identical logic. *Nature* **443**, 415–420 (2006).
15. S. Tan, T. J. Richmond, Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* **391**, 660–666 (1998).
16. E. Gocke, T. R. Manney, Expression of radiation-induced mutations at the arginine permease (CAN1) locus in saccharomyces cerevisiae. *Genetics* **91**, 53–66 (1979).
17. C. D. Aakre *et al.*, Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
18. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).
19. V. O. Pokusaeva *et al.*, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLOS Genet.* **15**, e1008079 (2019).
20. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
21. T. N. Starr, J. M. Flynn, P. Mishra, D. N. A. Bolon, J. W. Thornton, Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4453–4458 (2018).
22. O. Puchta *et al.*, Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840–844 (2016).
23. R. Brent, M. Ptashne, A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **43**, 729–736 (1985).
24. R. D. Kornberg, Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.* **30**, 235–239 (2005).
25. P. Tompa, M. Fuxreiter, Fuzzy complexes: Polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
26. L. M. Tuttle *et al.*, Gcn4-mediator specificity is mediated by a large and dynamic fuzzy protein-protein complex. *Cell Rep.* **22**, 3251–3264 (2018).
27. P. S. Brzovic *et al.*, The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. *Mol. Cell* **44**, 942–953 (2011).
28. A. L. Sanborn *et al.*, Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to mediator. *ELife* **10**, e68068 (2021).
29. L. Warfield, L. M. Tuttle, D. Pacheco, R. E. Klevit, S. Hahn, A sequence-specific transcription activator motif and powerful synthetic variants that bind mediator using a fuzzy protein interface. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3506–E3513 (2014).
30. R. D. Gietz, R. A. Woods "Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method" in *Methods in Enzymology*, C. Guthrie, G. R. Fink, Eds. (Elsevier, 2002), **vol. 350**, pp. 87–96.
31. L. Benatuil, J. M. Perez, J. Belk, C.-M. Hsieh, An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **23**, 155–159 (2010).
32. G. F. Sprague "Assay of yeast mating reaction" in *Methods in Enzymology,* C. Guthrie, G. R. Fink, Eds. (Elsevier, 1991), **vol. 194**, pp. 77–93.
33. K. R. Fowler, F. Leon, A. D. Johnson, Ancient transcriptional regulators can easily evolve new pair-wise cooperativity. NCBI Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE233191. Deposited 22 May 2023.