www.sciencemag.org/content/342/6154/104/suppl/DC1

Supplementary Materials for

Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution

Christopher R. Baker, Victor Hanson-Smith, Alexander D. Johnson*

*Corresponding author. E-mail: ajohnson@cgl.ucsf.edu

Published 4 October 2013, *Science* **342**, 104 (2013)
DOI: 10.1126/science.1240810

This PDF file includes:

Supplementary Materials

Material & Methods

Ancestral reconstruction of MADS-box domains

Orthologs of *S. cerevisiae MCM1* and *ARG80* were defined by a tBLASTN search of the 35 additional hemiascomyete genomes available at the NCBI website, as well as a BLASTp search of the collection of yeast genomes available on the Yeast Gene Order Browser webpage (*25*) (http://wolfe.gen.tcd.ie/ygob/) (Table S1). To eliminate false positives, hit sequences were reverse BLASTed against the *S. cerevisiae* genome and we eliminated any BLAST hits that did not have either *ARG80* or *MCM1* as their best match within the *S. cerevisiae*. A handful of phylogenetically non-informative sequences were excluded because of their near-identity conservation with sequences that were retained in our dataset. With the remaining 22 sequences, a multiple protein sequence alignment was inferred using PRANK with default settings (*26*). This alignment was best-fit by the Jones-Taylor-Thornton matrix (JTT) with gamma-distributed rate variation (+G), according to the Akaike Information Criterion as implemented in PROTTEST (*27*).

Using JTT+G, we used a maximum likelihood (ML) algorithm (*28*) to infer the ancestral amino acid sequences with the highest probability of producing all the extant sequence data. Specifically, we used PhyML v3.0 to infer the ML topology, branch lengths, and evolutionary rates (*29*). We optimized the topology using the best result from Nearest-Neighbor-Interchange and Subtree Pruning and Regrafting; we optimized branch lengths and all other model parameters using the default hill-climbing algorithm in PhyML. Statistical support for branches was calculated as approximate likelihood ratios.

We reconstructed ML ancestral states at each site for all ancestral nodes in our ML tree using Lazarus (*30*). We used *Y. lipolytica* as the outgroup sequence to the root tree, and placed ancestral insertion/deletion characters according to Fitch's parsimony. We characterized uncertainty for AncMADS, AncArg80, and AncMcm1 by binning the posterior probabilities of states into 5%-sized bins and counting the proportion of ancestral sites within each bin (Fig. S1). All sequence sites within the MADS-box domain in the AncMADS and AncMcm1 ancestors, and 79 out of 85 sites in AncArg80

MADS-box, were strongly supported with probability greater than 0.8 (Fig. S1, Table S2). We ignored the alternate states at these sites because each of these sites experienced degenerative substitutions (i.e. loss of a conserved amino acid identity) on the branch leading to AncArg80, and therefore, the specific amino acid identity of these sites are of minimal or no importance. Further, four of these substitutions occur on surfaces that are structurally located away from the interfaces for DNA-binding and cofactor interactions. Combined with the fact that dN/dS tests failed to identify any sites under positive selection—including these six sites—we interpret the uncertainty at these sites as a reflection of their degenerative evolutionary history.

Tests for positive selection

We tested for signals of positive selection in the evolution of paralog Arg80, using several models of synonymous and non-synonymous codon evolution implemented in PAML (*31*). We collected from GenBank nucleotide coding sequences for the Mcm1 and Arg80 proteins used in our study. We first estimated maximum likelihood values for a simple codon model (i.e. the "M0" model) with a single ratio (ω) of the nonsynonymous substitution rate (dN) to the synonymous substitution rate (dS). We tested for signals of positive selection on the branch leading to Anc.Arg80 by comparing the simple M0 model to more complex models that incorporate additional dN/dS ratios to allow for adaptive evolution and neutral evolution. Specifically, we estimated values for the so-called "branch" model that includes two free dN/dS ratios: ω_1 on the branch leading to AncArg80, and ω_2 on all other branches. We also estimated values for the so-called "branch-neutral" model that tests for neutrality on the branch model by fixing the value of ω_1 at 1.0. We compared the goodness-of-fit of the branch and branch-neutral models to the M0 model, using a likelihood ratio test as implemented in Excel (Table S3).

We next tested for signals of positive selection in specific sites. We estimated values for a "nearly-neutral" mixture model in which sites are classified into two categories: sites under purifying selection ($0 < dN/dS < 1$), and sites under neutral evolution ($dN/dS = 1$). We compared this to a more complex model of positive selection, which is identical to the nearly-neutral model, but with an additional dN/dS category for

sites evolving under positive selection (dN/dS > 1). We compared the goodness-of-fit of these two models using a likelihood ratio test (Table S3).

Homology modeling

The structures of AncMADS, AncMcm1, and AncArg80 were predicted by homology modeling, based on the crystal structure of *S. cerevisiae* Mcm1 bound to MATα2 on *STE6* operator DNA (*18*). We used Modeller software (*32*) to infer each ancestral structure 100 times, using default settings, and then chose the lower-energy iteration for these structures. The resultant models were then aligned and visually compared in MacPyMol [http://www.pymol.org].

RNA isolation & quantification

Strains were grown in YPD to OD_{600} = 0.8 and then centrifuged at 3000 x G for 5 minutes. The supernatant was removed and pellets were frozen in liquid nitrogen. RNA was isolated as previously described (*33*) with all volumes scaled appropriately. Total RNA was quantified by OD_{260} and its purity assessed using OD_{260} /OD_{230} and OD_{280} /OD_{260} ratios. NanoString quantification of transcript abundance was performed by NanoString Core facility in Seattle, Washington, USA.

For RT-qPCR, *S. cerevisiae* strains were grown in YPD to OD_{600} = 0.8 and then centrifuged at 3000 x G for 5 minutes. The supernatant was removed and pellets were then frozen in liquid nitrogen. RNA was isolated and reverse transcribed (using SuperScript II) as previously described (*33*) with all volumes scaled appropriately. cDNAs were quantified with a Bio-Rad CFX96 Real Time machine in a standard 25 µl reaction using Sybr green and primer sequences listed in Table S4.

Growth assays

For the ornithine growth assay, cells were grown overnight in YPD and then inoculated at an OD_{600} of 0.1 into a minimal ornithine growth media, which included only 2% glucose, yeast nitrogen bases (without ammonium), and 1 mg/ml ornithine as the sole nitrogen source for amino-acid production. For YPD growth assay, cells were grown overnight in YPD and then inoculated at an OD_{600} of 0.05 in YPD.

Gel shift experiments

The 82 amino-acid ancestral MADS-box domains were expressed in *E. coli* BL21 (DE3) cells from the pET26b plasmid. 100 ml of cells were induced at an OD_{600} of 0.6 with 1 mM IPTG and incubated overnight at 16 ^{o}C and 300 rpm. Cells were lysed, proteins purified using the His6x tag, and epitope tags removed from the recombinant protein as previously described (*34*). *S. cerevisiae ARG3 cis*-regulatory sequence oligonucleotide probes were labeled with P^{32} γ-ATP using T4 PNK. Binding conditions were 50 mM Tris [pH = 8], 100 mM NaCl, 10% Glycerol, 5 mM $MgCl_2$, 5mM β-mercaptoethanol, 50μg/mL Poly(dI-dC) (limits non-specific protein:DNA-binding), and 1.2 μM labeled oligonucleotide. Labeled DNA was incubated in the presence of various concentrations of ancestral MADS-box proteins for 30 minutes. After 30 minutes (time zero), 120 μM unlabeled oligonucleotide was added to the reaction. The amount of protein bound to P^{32}-labeled DNA was quantified by phosphoimaging. Following the introduction of saturating levels of unlabeled DNA, the dissociation of protein from labeled DNA can be modeled by an exponential decay curve. The values shown in Figure 3D are the natural log of the ratio of P^{32} bound DNA at the indicated time point to the amount of P^{32} bound DNA at time zero.

We assayed DNA-binding on a *cis*-regulatory sequence from the ARG activated gene *CAR2* that is ~150bp upstream of the translation start site. The decision to select this regulatory sequence was made based on three criteria. First, the stability of Arg80 binding at the ARG genes has been shown to be higher than at other MADS-box regulatory sites, such as those at the mating genes (*16*). Second, this site supports binding by both paralogs and when tested *in vitro,* it had behaved in a representative fashion relative to other ARG gene regulatory sites (*20*). Finally, this site closely-matched the consensus binding sequences for Mcm1 and Arg80 (*8, 20*).

Strain construction

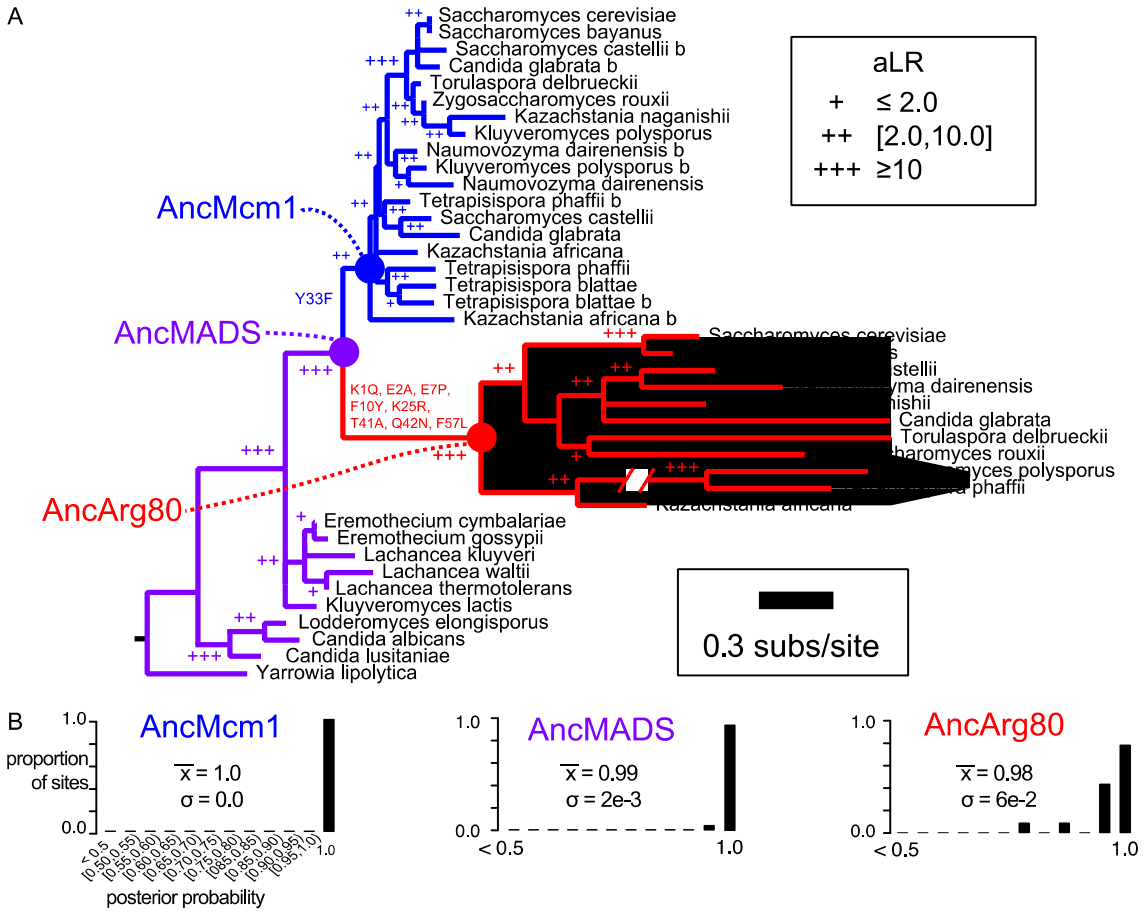
A list of all strains used in this study can be found in Table S5.

Deletion of *S. cerevisiae* MADS-box proteins was performed using a standard lithium acetate transformation and the pFA6a drug marker series (NATmx6, KANmx6)

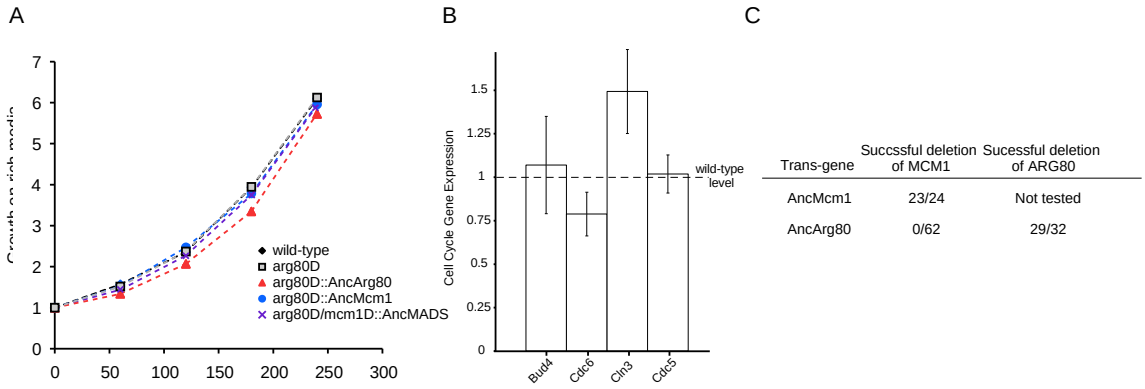
(*35*) with 60bp of homology targeting to the relevant MADS-box gene locus. Since the duplicates are in tandem, a single cassette was used to delete both proteins in some strains.

Ancestral Mcm1 MADS-box domains were synthesized and codon-optimized for expression in *S. cerevisiae* by DNA 2.0 (South San Francisco, CA). Ancestral sequences were then fused to the *S. cerevisiae* sequence flanking the endogenous MADS-box domain by fusion PCR (using the NEB Phusion polymerase and High-Fidelity buffer) and cloned into the pNH605 plasmid (for instance, *S. cerevisiae ARG80*-ancArg80 MADS-box-*S. cerevisiae ARG80* or *S. cerevisiae MCM1*-ancMADS MADS-box-*S. cerevisiae MCM1*). This flanking sequence extended beyond the coding sequence to include the full upstream region until the next ORF for endogenous expression level constructs and cloned into the pNH605 plasmid (using PspOMI/BamHI) or cloned directly into pNH605 vectors modified to include the TEF (pNH605-TEF) promoters (using BamHI/SacI) (*36*). All constructs were sequenced to check for mutations within the coding sequence and promoter. Once linearized (by cutting with PmeI), the pNH605 vector integrated into the *S. cerevisiae* genome at the *LUE2* locus in single copy.

S. cerevisiae strains were generated using a standard lithium acetate transformation in the W303 and for just the ornithine growth experiments in a S288c prototrophic and S288c *ura-* background.

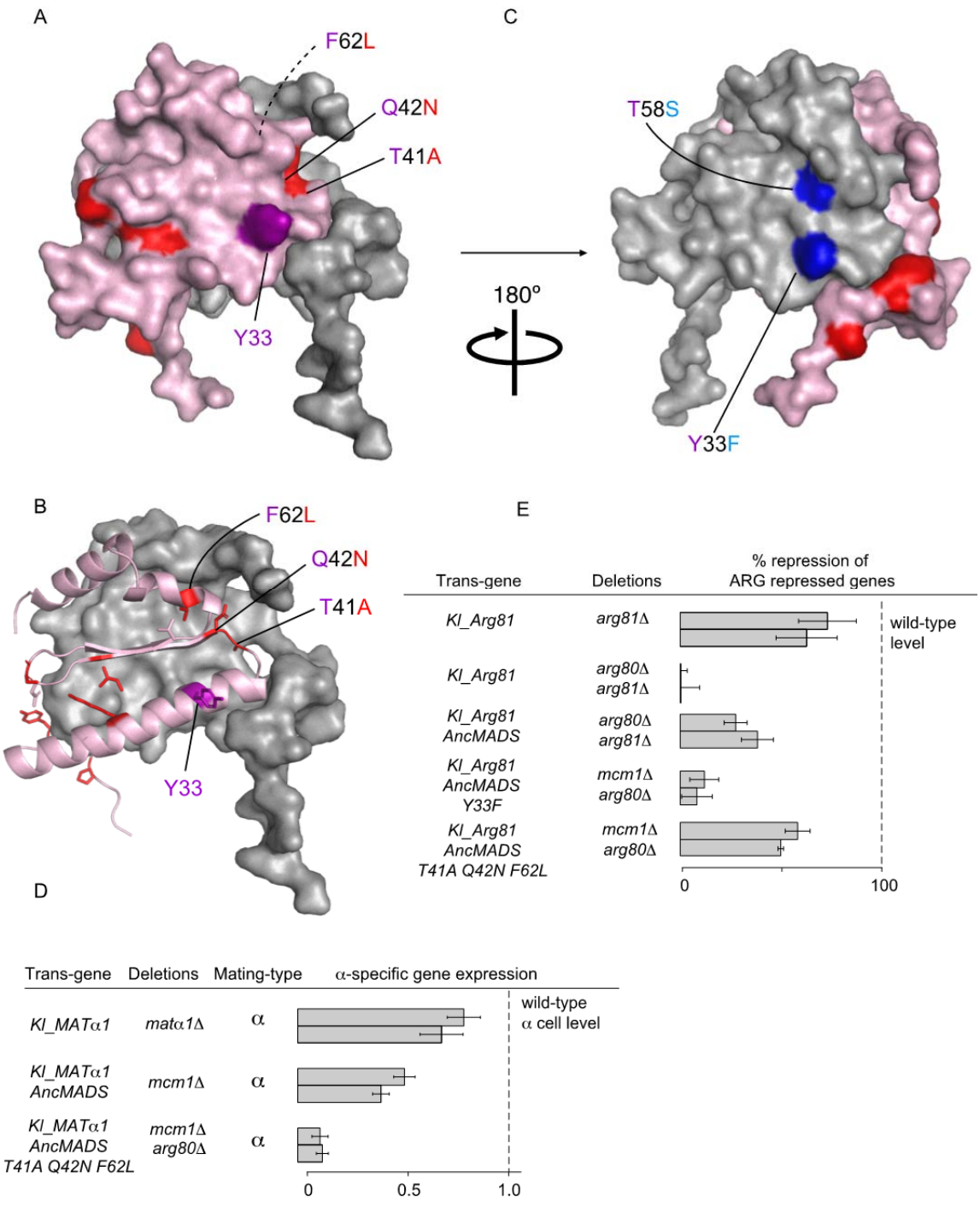


Supplemental Figure S1—Evolution of MADS-box proteins in hemiascomycete yeast. (A) Maximum likelihood phylogeny of full-length protein sequences for Mcm1 and Arg80 in hemiascomycete (*Saccharomycotina*) yeast species. Terminal taxa names correspond to sequences in Table S1. Branch lengths express average amino acid substitutions per site. Symbols on internal branches correspond to approximate likelihood ratios (aLR), expressing relative likelihood support for the monophyly of the descendant clade. Relevant MADS-box mutations are listed on the branch leading to ancestor Anc.Mcm1 (in blue) and to Anc.Arg80 (in red). (B) Support for reconstructions of ancestors AncMcm1, AncMADS, and AncArg80. Sites were binned into 5%-sized bins, based on the posterior probability of their reconstructed state, and then the proportion of sites in bins was counted. x (mean) and σ (standard deviation) of the posterior probabilities of the best amino acid state in the reconstructed ancestors. Note that an additional duplication of *ARG80* and *MCM1* occurred at the yeast whole genome duplication (*37, 38*). Many post-whole genome duplication species retain two copies of Mcm1 and thus, there are more representatives of Mcm1 than Arg80 in our MADS-box gene tree.



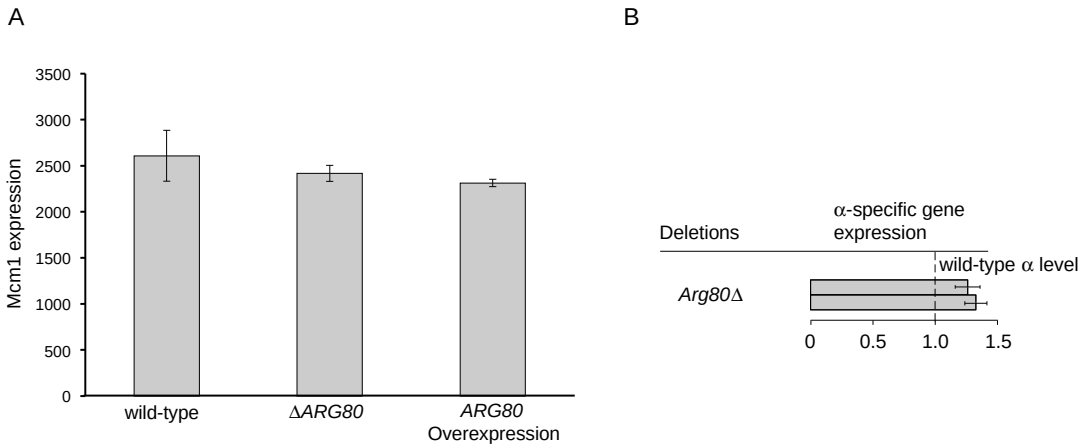
Trans-gene	Succssful deletion of MCM1	Sucessful deletion of ARG80
AncMcm1	23/24	Not tested
AncArg80	0/62	29/32

Supplemental Figure S2—The pre-duplication ancestral gene complements both paralogs. (A**)** Replacement of endogenous MADS-box proteins with AncMADS does not cause a log-phase growth-rate defect in rich-media (YPD). Mcm1 is a cell-cycle regulator and when compromised, growth is severely compromised or non-existent (*9*). The purple line marked by "x"s shows that the pre-duplication ancestral gene can supply the function of both modern paralogs. "Growth" on the y-axis is the ratio of OD_{600} divided by OD_{600} at time zero. **(**B**)** The pre-duplication AncMADS protein can complement both modern paralogs in *S. cerevisiae* at the cell-cycle genes. The endogenous MADS-box proteins were replaced by the pre-duplication AncMADS gene to create a *mcm1Δ/arg80Δ*::AncMADS strain. We then quantified the expression of the four listed MADS-box regulated cell cycle genes and displayed their expression relative to wild-type. Expression data was collected using NanoString. Mean and standard error were determined from 3 replicates. (C) AncArg80 does not complement the essential gene *MCM1* in *S. cerevisiae*. Strains with AncArg80 and AncMcm1 incorporated into the genome at the *LEU2* locus were transformed with kanamycin resistance gene with flanking sequences targeting either the *ARG80* or *MCM1* loci. Number of colonies screened is quantified in the denominator for each transformation.



Supplemental Figure S3— Divergence in cofactor-binding following gene duplication. (A-C**)** Structural homology model of AncMADS from *S. cerevisiae* Mcm1 structure from Tan & Richmond (*18*). The surface of AncMcm1 is modeled in grey and that of AncArg80 is modeled in pink. Residues that mutated between AncMcm1 and AncArg80 (relative to AncMADS) are labeled red and blue, respectively. **(**B**)** Same as **(**A**)**, but AncArg80 is represented as a cartoon. **(**C**)** Residues that mutated between

AncMADS and AncMcm1 are labeled in blue. **(**D-E**)** Gene expression profiling reveals that mutations in the pre-duplication AncMADS protein in *S. cerevisiae* compromises interaction with *K. lactis* cofactors. It was important for us to rule out the possibility that the gene expression effects we observed in Figure 3B were specific to the interactions between the MADS-box proteins and the *S. cerevisiae* cofactors Arg81 and Matα1. To test this end, we opted to replace the endogenous Arg81 and Matα1 with the orthologous proteins from a species that branches before the MADS-box duplication event, *K. lactis*. Consistent with our findings in Figure 3B, we found that the post-duplication changes to the ancestral Arg81/Matα1 interaction surface also selectively compromised the interaction with *K. lactis* α1 and Arg81. Gene expression quantified using NanoString. **(**E**)** MADS-box activated mating genes (α-specific genes). Columns: 1- *SAG1,* 2- *MFα1*. **(**F**)** Columns: 1- *ARG3,* 2- *ARG5,6.* Mean and standard error determined from 3 replicates.



Supplemental Figure S4— Paralog interference between Arg80 and Mcm1. (A**)** We hypothesized that Arg80 might indirectly regulate this gene set through repression of *MCM1* expression. Inconsistent with this hypothesis, *MCM1* expression remained constant when the expression of *ARG80* was changed. Expression data was collected using NanoString. Mean and standard error were determined using 3 to 5 replicates. **(**B**)** Expression data collected using NanoString for α-specific genes (*MFα1* and *SAG1*) confirms that deletion of *ARG80* increases expression level. Mean and standard error were determined from 5 replicates.

Supplemental Table S1— Aligned MADS-box domain protein sequences from Hemiascomycetes used in this study.

>Kluyveromyces_(Vanderwaltozyma) polysporus Arg80 syntenic ortholog
GKKRKLQLKYISNKSRRQVTFSKRRIGLMKKCYELSVMTGVNMLLLVASDSKLVYTFSTPKLRRFIEENEGKSLVRNCLKKED
>Tetrapisispora phaffi Arg80 syntenic ortholog
TKNKKIKIDFIVNKTKRRLTYNTRRVGLMKKSYELSLITGVNILLLISSNDDYVYFFSSPKLRNFVNCNEGQKLIRKCLQEES
>Candida glabrata Arg80 syntenic ortholog
GKRRKAPLKYIENTTRRHVTFAKRRHGIMKKAYELAVMTGANVLLLILSPKGLVYTFATPSLQPLIRDDPGKELIRRCLNQDN
>Kazachstania africana Arg80 syntenic ortholog
RARRKNPIKYIENKTRRQVTFAKRRHGLMKKAYELSVMTGANILLLIVSNTGLVYTFTTPKLEPVVINEEGKNLIRACLNAED
>Candida glabrata Mcm1 syntenic ortholog
KDRRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLIVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Tetrapisispora blattae Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Tetrapisispora phaffi Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Kluyveromyces (Vanderwaltozyma) polysporus Mcm1 whole genome duplicate
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Candida albicans Mcm1
KERRKIEIKFIQEKSRRHITFSKRKAGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKLQPLVTKSEGKNLIQACLNAPE
>Yarrowia lipolytica Mcm1
RERRKIEIKFIQDKSRRHITFSKRKAGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKLQPLVTKPEGKNLIQACLNASD
>Tetrapisispora blattae Mcm1 whole genome duplicate
RERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Tetrapisispora phaffi Mcm1 whole genome duplicate
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Kazachstania africana Mcm1 whole genome duplicate
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Candida lusitaniae Mcm1
KERRKIEIKFIQDKSRRHITFSKRKAGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKLQPLVTKSEGKNLIQACLNAPE
>Lodderomyces elongisporus Mcm1
KERRKIEIKFIQDKSRRHITFSKRKAGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKLQPLVTKPEGKNLIQACLNAPE
>Kazachstania naganishii Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Zygosaccharomyces rouxii Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Naumovozyma dairenensis Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Torulaspora delbrueckii Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Saccharomyces (Naumovozyma) castellii Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLIVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Candida glabrata Mcm1 syntenic ortholog 2
RERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Kazachstania africana Mcm1 syntenic ortholog
KERKKIDIKYIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQPEGRNLIQACLSAPD
>Sacchoromyces cerevisiae Mcm1
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Saccharomyces bayanus Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Naumovozyma dairenensis Mcm1 syntenic ortholog 2
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNVPD

>Saccharomyces (Naumovozyma) castellii Mcm1 whole genome duplicate
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Kluyveromyces (Vanderwaltozyma) polysporus Mcm1 syntenic ortholog
KERRKIEIKFIENKTRRHVTFSKRKHGIMKKAFELSVLTGTQVLLLVVSETGLVYTFSTPKFEPIVTQQEGRNLIQACLNAPD
>Lachancea kluyveri Mcm1
KERRKIEIKFIQDKTRRHITFSKRKHGIMKKAYELSVLTGTQVLLLIVSETGLVYTYTTPKFQPIVKEPEGRNLIQACLNAPD
>Kluyveromyces lactis Mcm1
KERRKIEIKFIQDKTRRHITFSKRKHGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKFQPIVTQPEGKNLIQACLNAPE
>Eremothecium gossypii Mcm1
KERRKIEIKFIQDKTRRHITFSKRKHGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKFQPIVTQPEGKNLIQACLNAPE
>Eremothecium cymbalariae Mcm1
KERRKIEIKFIQDKTRRHITFSKRKHGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKFQPIVTQPEGKNLIQACLNAPE
>Lachancea waltii Mcm1
KERRKIEIKFIQDKTRRHITFSKRKHGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKFQPIVTQPEGKNLSRRVSTR--
>Lachancea waltii Mcm1 2
KERRKIEIKFIQDKTRRHITFSKRKHGIMKKAYELSVLTGTQVLLLVVSETGLVYTFTTPKFQPIVTQPEGKNLIQACLNAPE
>Saccharomyces cerevisiae Arg80 syntenic ortholog
VTRRKQPIRYIENKTRRHVTFSKRRHGIMKKAYELSVLTGANILLLILANSGLVYTFTTPKLEPVVREDEGKSLIRACINASD
>Torulaspora delbrueckii Arg80 syntenic ortholog
GAKQKVPVKYIANPARRHVTFAKRRHGIMKKAYELSVLTGANVLLLILSKSGLVYTFTTPKLERIVREQEGKRLIRQCLQEET
>Kazachstania naganishii Arg80 syntenic ortholog
DVRRKVPIKYLENRTRRQVTFAKRRHGIMKKAYELSVLTGSNILLLILSRSGLVYTFSTPKLEPIIRDEEGKTLIRKCLNAGE
>Saccharomyes bayanus syntenic ortholog
TARQKQPIKYIENKTRRHVTFSKRRHGIMKKAYELSVLTGANILLLILANSGLVYTFTTPKLEPLVREDEGKNLIKACINAPD
>Saccharomyces (Naumovozyma) castellii Arg80 syntenic ortholog
SMRQRIPIKYIENKTRRHVTFAKRRHGIMKKAYELSVLTGANVLLLILSGSGLVYTFSTPKLEPVIRDEEGKGLIRACLNGPD
>Naumovozyma dairenensis Arg80 syntenic ortholog
SSKQKIPISFIENKSRRHVTFAKRRHGIMKKAYELSVLTGANVLLLILSSSGLVYTFTTPKLEPVIRDDEGKNLIRKCLGAPD
>Zygosaccharomyces rouxii Arg80 syntenic ortholog
QQKRKYPIKYIENRTRRHVTFAKRRHGIMKKAYELSVLTGANVLLLILSNNGLVYTFTTPKLEPVVREREGKELVWKCLDGRP

Supplemental Table S2 – Posterior Probability Distributions for Reconstructed Ancestral Proteins. Site numbers correspond to MADS-box domains sites in the multiple sequence alignment inferred by PRANK. Decimals to the right of amino acids express the posterior probability of that amino acid at the corresponding site. Distributions are highlighted by ancestor: AncMADS in light purple, AncMcm1 in light blue, and AncArg80 in light pink.

Site	Anc.MADS		Anc.Mcm1			Anc.Arg80		
	state	PP	state	PP	Alternates	state	PP	Alternates
1	K	1	K	1		Q	0.118	*T 0.109 S 0.100 N 0.091 G 0.090 A 0.082 R 0.074 K 0.072 E 0.057 P 0.048 D 0.043 V 0.031 H 0.027 I 0.022 L 0.016 M 0.012 Y 0.004 C 0.003 F 0.002*
2	E	1	E	1		A	0.244	*T 0.16 V 0.152 I 0.089 S 0.079 M 0.045 P 0.041 N 0.033 G 0.032 L 0.031 Q 0.029 E 0.018 D 0.014 K 0.012 R 0.008 H 0.006 F 0.003 Y 0.002 C 0.001*
3	R	1	R	1		R	1	
4	R	1	R	1		R	1	
5	K	1	K	1		K	1	
6	I	1	I	1		I	0.818	*V 0.168 L 0.004 N 0.004 M 0.002 T 0.002 Q 0.001*
7	E	1	E	1		P	0.995	*Q 0.004*
8	I	1	I	1		I	1	
9	K	1	K	1		K	1	
10	F	1	F	1		Y	0.997	*F 0.003*
11	I	1	I	1		I	1	
12	Q	1	E	1		E	1	
13	D	1	N	1		N	1	
14	K	1	K	1		K	0.999	*R 0.001*
15	T	1	T	1		T	1	
16	R	1	R	1		R	1	
17	R	1	R	1		R	1	

18	H	1	H	1		H	0.986	*Q 0.013*
19	I	1	V	1		V	0.999	*I 0.001*
20	T	1	T	1		T	1	
21	F	1	F	1		F	1	
22	S	1	S	1		S	1	
23	K	1	K	1		K	1	
24	R	1	R	1		R	1	
25	K	1	K	1		R	0.999	*K 0.001*
26	H	1	H	1		H	1	
27	G	1	G	1		G	1	
28	I	1	I	1		I	0.996	*L 0.003*
29	M	1	M	1		M	1	
30	K	1	K	1		K	1	
31	K	1	K	1		K	1	
32	A	1	A	1		A	1	
33	Y	1	F	1		Y	1	
34	E	1	E	1		E	1	
35	L	1	L	1		L	1	
36	S	1	S	1		S	1	
37	V	1	V	1		V	1	
38	L	1	L	1		L	0.982	*M 0.018*
39	T	1	T	1		T	1	
40	G	1	G	1		G	1	
41	T	1	T	1		A	0.993	*T 0.006*
42	Q	1	Q	1		N	1	
43	V	1	V	1		V	0.506	*I 0.494*
44	L	1	L	1		L	1	
45	L	1	L	1		L	1	
46	L	1	L	1		L	1	
47	V	1	V	1		I	0.985	*V 0.015*
48	V	1	V	1		V	0.936	*L 0.057 I 0.006 M 0.001*
49	S	1	S	1		S	1	
50	E	1	E	1		N	0.988	*D 0.009 E 0.001 K 0.001 S 0.001*
51	T	1	T	1		T	0.707	*S 0.291 A 0.001 N 0.001*

52	G	1	G	1		G	1	
53	L	1	L	1		L	1	
54	V	1	V	1		V	1	
55	Y	1	Y	1		Y	1	
56	T	1	T	1		T	1	
57	F	1	F	1		F	1	
58	T	1	S	0.989	T 0.011	T	0.994	*S 0.006*
59	T	1	T	1		T	1	
60	P	1	P	1		P	1	
61	K	1	K	1		K	1	
62	F	1	F	1		L	0.997	*F 0.003*
63	E	1	E	1		E	0.999	
64	P	1	P	1		P	1	
65	I	1	I	1		V	0.962	*I 0.037*
66	V	1	V	1		V	0.999	*I 0.001*
67	T	1	T	1		T	0.455	*R 0.345 I 0.147 K 0.012 S 0.010 A 0.006 N 0.006 M 0.005 Q 0.005 V 0.005 E 0.001 H 0.001 L 0.001*
68	Q	1	Q	1		E	0.766	*D 0.1 Q 0.086 N 0.041 K 0.004 H 0.001*
69	P	1	Q	0.989	P 0.011	E	0.872	*D 0.087 Q 0.026 N 0.005 K 0.003 G 0.002 A 0.001 H 0.001 P 0.001 S 0.001 R 0.001 T 0.001*
70	E	1	E	1		E	1	
71	G	1	G	1		G	1	
72	K	1	R	1		K	0.999	*R 0.001*
73	N	1	N	1		N	0.999	
74	L	1	L	1		L	1	
75	I	1	I	1		I	1	
76	Q	1	Q	1		R	0.997	*Q 0.002 K 0.001*
77	A	1	A	1		A	0.999	
78	C	1	C	1		C	1	
79	L	1	L	1		L	1	

80	N	1	N	1		N	1	
81	A	1	A	1		A	1	
82	P	1	P	1		P	0.995	*Q 0.003 A 0.001 E 0.001 S 0.001*
83	E	1	D	1		D	1	

Model Name	Fixed Params.	Free Parameters, Estimated with Maximum Likelihood	Log ML	N free params.	LRT	Significant?
M0	none	k=1.31, w=0.00547	-4155.567	2	n/a	
Branch	None	k=1.30933, w=0.00527, w branch=179.82947	-4155.483	3	LTR(M0): P = 0.68	No
Branch-Neutral	w_branch = 1.0	k=1.30990, w=0.00534	-4155.489	2	LTR(M0): P = 0.69	No
Sites	w2=1.0	k=1.30332, w1=0.00546, p(w1)=0.99999	-4148.058	3	n/a	
Sites-Neutral	W2=1.0	k=1.30382, w1=0.00545, w3=1.000001, p(w1)=0.99999, p(w2)=0.0	-4148.058	5	LTR(Sites): P = 0.99	No

Supplemental Table S3— Tests for Positive Selection in the Evolution of *ARG80*.
We tested for signals of positive selection using several codon models that include various combinations of free and fixed parameters (see Methods and Materials). We computed the log likelihood (**Log ML)** of the codon sequence alignment evolving according to the values of the fixed and free parameters. We tested for goodness-of-fit of models by using a likelihood ratio test **(LRT)** to compare complex models to simpler models. □, the ratio of of nonsynonymous (dN) to synonymous (dS) codon substitutions rates. **k**, the ratio of nucleotide transitions to transversions. **p**, the probability of sites evolving according to the corresponding □ ratio. Note that for the sites-neutral model, the ML value of w3 is indeed 1.000001, effectively collapsing the sites-neutral model into the simpler sites model. It should be noted that the underlying rates of synonymous substitutions (i.e. dS) often exceeded 2.0, indicating that dS is saturated and the power of this test to detect selection may be compromised.

Supplemental Table S4— qPCR Primers used in this study

URA6 set 1	ACGCCAAGGAGCTGTCATAC
URA6 set 1	TTCGAGCCCATGAACTTTCT
URA6 set 2	GACCGGTCGAAGAAGAATGA
URA6 set 2	GACGTGATTTGCGAGTGGTA
SAG1 set 1	GATGGACAGGCGCTATGAAT
SAG1 set 1	TGGAACAGCAGCAGTATTCG
SAG1 set 2	GGTACAGCTAGCGCCAAAAG
SAG1 set 2	TCACCACATGGCTGATCACT

Supplemental Table S5— Strain List

	Figures	Genotype	Origin
1	2b-e, 3b, S2a-b, S3e-f, S4a-b	W303 α cell *ura3-1 trp1-1 can1-100 ade2-1 lue2-3, 112*	Standard strain
2	2d-e	W303 **a** cell *ura3-1 trp1-1 can1-100 ade2-1 lue2-3, 112*	Standard strain
3	2b-c, 4d, S2a, S4a	W303 α cell *arg80*Δ::*NATmx6*	This work
4	2B, S2A	s288c prototrophic α	Standard strain
5	2B, S2A	s288c prototophic α *arg80*Δ::*NATmx6*	This work
6	2B, S2A	s288c prototrophic α *arg80*Δ::*NATmx6 mcm1*Δ::*Mcm1-KanMx6*	This work
7	2B, S2A	s288c prototrophic α *arg80*Δ::*NATmx6 mcm1*Δ::*ScMcm1(amino acids 1-15)-AncMADS-ScMcm1(amino acids 99-286)-KanMx6*	This work
8	2B, S2A	s288c *ura3-* α *arg80Δ::NATmx6 pRS316-arg80 promoter-ScArg80 (amino-acids 1-77)-AncArg80-(amino-acids 161-177)*	This work
9	2b-c, 4c	W303 a *arg80*Δ::NATmx6 *pNH605-arg80 promoter-ScArg80 (amino-acids 1-77)-AncArg80-(amino-acids 161-177)*	This work
10	4c, S4a	W303 α *arg80*Δ::NATmx6 *pNH605-TEF promoter arg80 promoter-ScArg80 (amino-acids 1-77)-AncArg80-(amino-acids 161-177)*	This work
11	numerous	W303 α *arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS-ScMcm1(amino acids 99-286)*	This work
12	2d-e	W303 α *mcm1*Δ::*KANMx6 pNH605-scMcm1*	This work
13	3b	W303 α *arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS Y33F -ScMcm1(amino acids 99-286)*	This work
14	3b	W303 α *arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS T41A Q42N F62L -ScMcm1(amino acids 99-286)*	This work
15	4C	W303 α *Arg80Δ::NATmx6* pNH605-A*rg80 promoter-AncArg80 Q1K A2E P7E Y10F R25K*	This work
16	4C	W303 α *arg80Δ::NATmx6* pNH605-*Arg80 promoter-AncArg80 Q1K A2E P7E Y10F R25K*	This work
17	S3e	W303 α Sc*_arg81*Δ::*Kl_ARG81-URA3*	This work
18	S3e	*W303* α *arg80*Δ::*NATMx6*, *Sc_arg81*Δ::*Kl_ARG81-URA3*	This work
19	S3e	W303 α *arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS-ScMcm1(amino acids 99-286) Sc_arg81*Δ::*Kl_ARG81-URA3*	This work
20	S3e	W303 α *arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS Y33F-ScMcm1(amino acids 99-286) Sc_arg81*Δ::*Kl_ARG81-URA3*	This work
21	S3e	W303 α *arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS T41A Q42N F62L-ScMcm1(amino acids 99-286) Sc_arg81*Δ::*Kl_ARG81-URA3*	This work
22	S3f	W303 mat*Δ* p414-TEF *Kl_MATα1*	This work
23	S3f	W303 mat*Δ* arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS-ScMcm1(amino acids 99-286) p414-TEF Kl_MATα1	This work
24	S3f	W303 mat*Δ* arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS *Y33F* -ScMcm1(amino acids 99-286) p414-TEF Kl_MATα1	This work
25	S3f	W303 mat*Δ* arg80Δ/mcm1Δ::KANmx6 pNH605 ScMcm1(amino acids 1-15)-AncMADS T41A Q42N F62L-ScMcm1(amino acids 99-286) p414-TEF Kl_MATα1	This work

References and Notes

1. A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, J. Postlethwait, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999). Medline

2. H. Innan, F. Kondrashov, The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010). Medline doi:10.1038/nrg2689

3. A. Wagner, Selection and gene duplication: A view from the genome. *Genome Biol.* **3**, reviews1012.1–reviews1012.3 (2002). Medline doi:10.1186/gb-2002-3-5-reviews1012

4. A. Wagner, How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.* **270**, 457–466 (2003). Medline doi:10.1098/rspb.2002.2269

5. J. T. Bridgham, J. E. Brown, A. Rodríguez-Marí, J. M. Catchen, J. W. Thornton, Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet.* **4**, e1000191 (2008). Medline doi:10.1371/journal.pgen.1000191

6. F. Messenguy, E. Dubois, Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene* **316**, 1–21 (2003). Medline doi:10.1016/S0378-1119(03)00747-9

7. C. Boonchird, F. Messenguy, E. Dubois, Characterization of the yeast *ARG5,6* gene: Determination of the nucleotide sequence, analysis of the control region and of *ARG5,6* transcript. *Mol. Gen. Genet.* **226**, 154–166 (1991). Medline doi:10.1007/BF00273599

8. B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li, A. D. Johnson, The evolution of combinatorial gene regulation in fungi. *PLoS Biol.* **6**, e38 (2008). Medline doi:10.1371/journal.pbio.0060038

9. F. Messenguy, E. Dubois, Genetic evidence for a role for MCM1 in the regulation of arginine metabolism in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13**, 2586–2592 (1993). Medline

10. A. Bender, G. F. Sprague Jr., MATα1 protein, a yeast transcription activator, binds synergistically with a second protein to a set of cell-type-specific genes. *Cell* **50**, 681–691 (1987). Medline doi:10.1016/0092-8674(87)90326-6

11. A. E. Tsong, M. G. Miller, R. M. Raisner, A. D. Johnson, Evolution of a combinatorial transcriptional circuit: A case study in yeasts. *Cell* **115**, 389–399 (2003). Medline doi:10.1016/S0092-8674(03)00885-7

12. C. R. Baker, B. B. Tuch, A. D. Johnson, Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7493–7498 (2011). Medline doi:10.1073/pnas.1019177108

13. M. J. Harms, J. W. Thornton, Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **20**, 360–366 (2010). Medline doi:10.1016/j.sbi.2010.03.005

14. T. B. Acton, J. Mead, A. M. Steiner, A. K. Vershon, Scanning mutagenesis of Mcm1: Residues required for DNA binding, DNA bending, and transcriptional activation by a

MADS-box protein. *Mol. Cell. Biol.* **20**, 1–11 (2000). Medline doi:10.1128/MCB.20.1.1-11.2000

15. N. Amar, F. Messenguy, M. El Bakkoury, E. Dubois, ArgRII, a component of the ArgR-Mcm1 complex involved in the control of arginine metabolism in *Saccharomyces cerevisiae*, is the sensor of arginine. *Mol. Cell. Biol.* **20**, 2087–2097 (2000). Medline doi:10.1128/MCB.20.6.2087-2097.2000

16. A. Jamai, E. Dubois, A. K. Vershon, F. Messenguy, Swapping functional specificity of a MADS box protein: Residues required for Arg80 regulation of arginine metabolism. *Mol. Cell. Biol.* **22**, 5741–5752 (2002). Medline doi:10.1128/MCB.22.16.5741-5752.2002

17. J. Mead, A. R. Bruning, M. K. Gill, A. M. Steiner, T. B. Acton, A. K. Vershon, Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. *Mol. Cell. Biol.* **22**, 4607–4621 (2002). Medline doi:10.1128/MCB.22.13.4607-4621.2002

18. S. Tan, T. J. Richmond, Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* **391**, 660–666 (1998). Medline doi:10.1038/35563

19. T. E. Hayes, P. Sengupta, B. H. Cochran, The human c-fos serum response factor and the yeast factors GRM/PRTF have related DNA-binding specificities. *Genes Dev.* **2**, 1713–1722 (1988). Medline doi:10.1101/gad.2.12b.1713

20. F. Messenguy, E. Dubois, C. Boonchird, Determination of the DNA-binding sequences of ARGR proteins to arginine anabolic and catabolic promoters. *Mol. Cell. Biol.* **11**, 2852–2863 (1991). Medline

21. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012). Medline

22. M. Lynch, J. S. Conery, The origins of genome complexity. *Science* **302**, 1401–1404 (2003). doi:10.1126/science.1089370

23. M. Y. Dennis, X. Nuttle, P. H. Sudmant, F. Antonacci, T. A. Graves, M. Nefedov, J. A. Rosenfeld, S. Sajjadian, M. Malig, H. Kotkiewicz, C. J. Curry, S. Shafer, L. G. Shaffer, P. J. de Jong, R. K. Wilson, E. E. Eichler, Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012). Medline doi:10.1016/j.cell.2012.03.033

24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

25. J. L. Gordon, D. Armisén, E. Proux-Wéra, S. S. ÓhÉigeartaigh, K. P. Byrne, K. H. Wolfe, Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20024–20029 (2011). Medline doi:10.1073/pnas.1112808108

26. A. Löytynoja, N. Goldman, An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10557–10562 (2005). Medline doi:10.1073/pnas.0409137102

27. F. Abascal, R. Zardoya, D. Posada, ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005). Medline doi:10.1093/bioinformatics/bti263

28. Z. Yang, S. Kumar, M. Nei, A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650 (1995). Medline

29. S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003). Medline doi:10.1080/10635150390235520

30. V. Hanson-Smith, B. Kolaczkowski, J. W. Thornton, Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol. Biol. Evol.* **27**, 1988–1999 (2010). Medline doi:10.1093/molbev/msq081

31. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). Medline doi:10.1093/molbev/msm088

32. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali , Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **2**, 2.9.1–2.9.31 (2007).

33. Q. M. Mitrovich, B. B. Tuch, C. Guthrie, A. D. Johnson, Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*. *Genome Res.* **17**, 492–502 (2007). Medline doi:10.1101/gr.6111907

34. M. B. Lohse, R. E. Zordan, C. W. Cain, A. D. Johnson, Distinct class of DNA-binding domains is exemplified by a master regulator of phenotypic switching in *Candida albicans*. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14105–14110 (2010). Medline doi:10.1073/pnas.1005911107

35. M. C. Lorenz, R. S. Muir, E. Lim, J. McElver, S. C. Weber, J. Heitman, Gene disruption with PCR products in *Saccharomyces cerevisiae*. *Gene* **158**, 113–117 (1995). Medline doi:10.1016/0378-1119(95)00144-U

36. A. H. Chau, J. M. Walter, J. Gerardin, C. Tang, W. A. Lim, Designing synthetic regulatory networks capable of self-organizing cell polarization. *Cell* **151**, 320–332 (2012). Medline doi:10.1016/j.cell.2012.08.040

37. K. H. Wolfe, D. C. Shields, Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997). Medline doi:10.1038/42711

38. M. Kellis, B. W. Birren, E. S. Lander, Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624 (2004). Medline doi:10.1038/nature02424