

Comprehensive Analysis of the *SUL1* Promoter of *Saccharomyces cerevisiae*

Matthew S. Rich,^{*,†} Celia Payen,^{*} Alan F. Rubin,^{*,†,††,***} Giang T. Ong,^{*} Monica R. Sanchez,^{*}

Nozomu Yachie,^{†,††,§§} Maitreya J. Dunham,^{*} and Stanley Fields^{*,§,***,1}

^{*}Department of Genome Sciences, [§]Department of Medicine, and ^{***}Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, [†]Synthetic Biology Division, Research Center for Advanced Science and Technology, the University of Tokyo, and ^{§§}PRESTO, Japan Science and Technology Agency, Meguro-ku, Tokyo 153-8904, Japan, [‡]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, Victoria 3052, Australia, ^{**}Department of Medical Biology, University of Melbourne, Victoria 3010, Australia, ^{††}Bioinformatics and Cancer Genomics Laboratory, Peter MacCallum Cancer Centre, Victoria 3002, Australia, and ^{†††}Institute for Advanced Bioscience, Keio University, Tsuruoka, Yamagata 997-0035, Japan

ORCID ID: 0000-0001-9944-2666 (M.J.D.)

ABSTRACT In the yeast *Saccharomyces cerevisiae*, beneficial mutations selected during sulfate-limited growth are typically amplifications of the *SUL1* gene, which encodes the high-affinity sulfate transporter, resulting in fitness increases of >35%. *Cis*-regulatory mutations have not been observed at this locus; however, it is not clear whether this absence is due to a low mutation rate such that these mutations do not arise, or they arise but have limited fitness effects relative to those of amplification. To address this question directly, we assayed the fitness effects of nearly all possible point mutations in a 493-base segment of the gene's promoter through mutagenesis and selection. While most mutations were either neutral or detrimental during sulfate-limited growth, eight mutations increased fitness >5% and as much as 9.4%. Combinations of these beneficial mutations increased fitness only up to 11%. Thus, in the case of *SUL1*, promoter mutations could not induce a fitness increase similar to that of gene amplification. Using these data, we identified functionally important regions of the *SUL1* promoter and analyzed three sites that correspond to potential binding sites for the transcription factors *Met32* and *Cbf1*. Mutations that create new *Met32*- or *Cbf1*-binding sites also increased fitness. Some mutations in the untranslated region of the *SUL1* transcript decreased fitness, likely due to the formation of inhibitory upstream open reading frames. Our methodology—saturation mutagenesis, chemostat selection, and DNA sequencing to track variants—should be a broadly applicable approach.

KEYWORDS high-throughput sequencing; transcription; promoter mutagenesis; gene amplification; chemostat; yeast; *Saccharomyces cerevisiae*

CHANGES in the extent or timing of gene expression can have profound effects on molecular and organismal phenotypes and thereby drive evolution (King and Wilson 1975; Wray 2007). Heritable noncoding variation can alter gene expression in *cis* or in *trans* (Skelly *et al.* 2009), and both have been shown to contribute significantly to gene expression variation (Ronald *et al.* 2005; Tirosh *et al.* 2009; Skelly

et al. 2011). Two primary mechanisms by which *cis* variation can increase gene expression are increases in gene copy number and point mutations in regulatory regions. However, the relative effect size of amplification compared to point mutation is not known.

It has been proposed that amplification is both quickly achieved and then reverts after fixation of fitness-increasing point mutations (Hendrickson *et al.* 2002; Yona *et al.* 2012). This mechanism raises the question of whether amplification is simply a transitional state that provides an increased chance for a beneficial point mutation to occur. Alternatively, gene amplification may be so frequently observed because the fitness effects it confers are greater than those achievable by point mutations. Gene amplifications have been found to be advantageous in many contexts, including phenotypic

Copyright © 2016 by the Genetics Society of America
doi: 10.1534/genetics.116.188037

Manuscript received January 27, 2016; accepted for publication February 21, 2016; published Early Online March 1, 2016.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.188037/-/DC1.

¹Corresponding author: Howard Hughes Medical Institute, Department of Genome Sciences, University of Washington, Box 355065, Seattle, WA 98195.
E-mail: fields@uw.edu

evolution (Hoekstra and Coyne 2007; Stern and Orgogozo 2008) and cancers (Brison 1993; Lengauer *et al.* 1998), but few experiments have addressed whether *cis*-regulatory mutations could also lead to similar effects. In one study, random mutations were introduced into yeast and assayed for their effects on the expression of a fluorescent reporter (Gruber *et al.* 2012). Strains with altered reporter expression were classified as having either *trans* or *cis* mutations, the *cis* mutations being either noncoding point mutations in the reporter construct or amplifications. As expected, amplifications increased reporter expression, but there was no difference in effect between these and fluorescence-increasing *cis*-regulatory point mutations (although only two such mutations were isolated). No strains with a reporter copy number greater than two were isolated, leaving open the possibility that higher-copy-number gene amplifications could enhance expression beyond that achieved by point mutation.

To study these questions directly, we turned to another case in which gene amplifications are adaptive. When *Saccharomyces cerevisiae* is subjected to long-term growth under sulfate limitation, a region of chromosome II containing the sulfate transporter gene *SUL1* is recurrently amplified after ~50 generations (Gresham *et al.* 2008; Payen *et al.* 2014). The amplification of this gene leads to a fitness increase of 37–51% (Gresham *et al.* 2008; Payen *et al.* 2014), depending on amplicon size and copy number. Amplification appears to be the preferred means by which fitness can be increased, as the appearance of coding or noncoding mutations at the *SUL1* locus is very rare. However, this preference for amplification of *SUL1* could have other possible explanations in addition to superior fitness, such as differences in the mutation rate between point mutations and amplifications. We sought to determine whether *cis*-regulatory mutations were capable of comparable effects by directly creating such mutations via mutagenesis, an approach that allowed us to skirt the potential effect of mutation rate.

The specifics of *SUL1*'s transcriptional regulation are largely unknown. *SUL1* is one of the 45 genes composing the core *Met4* regulon of genes involved in yeast sulfur metabolism (Lee *et al.* 2010). *Met4* is a transcriptional activator, but has no DNA-binding activity. It is targeted to promoter sequences by combinations of three transcription factors—*Cbf1*, *Met31*, and *Met32*—that themselves lack transcriptional activation activity. *Cbf1*, *Met31*, and *Met32* can induce transcription individually and in combination (Lee *et al.* 2010; Petti *et al.* 2012). *SUL1* is unannotated in many functional genomic studies on regulation, including RNA sequencing (Nagalakshmi *et al.* 2008), DNase I mapping (Hesselberth *et al.* 2009), and ChIP-Exo (Rhee and Pugh 2012), presumably due to its low expression in the sulfate-rich conditions used by these studies.

Mutagenesis has long been applied to study the function of noncoding sequence. Early work in yeast established the structure of eukaryotic promoters, delineating upstream activating sequences as the binding sites for transcriptional

regulators. These experiments changed promoter sequences by truncation, internal deletion, replacement of activating sequences, rearrangement, and random mutagenesis (Struhl 1989). However, these experiments were low throughput, assaying promoter variants one at a time. More recent studies that couple mutagenesis with high-throughput sequencing have increased the resolution and throughput of the analysis of eukaryotic *cis*-regulatory elements (Kwasnieski *et al.* 2012; Melnikov *et al.* 2012; Patwardhan *et al.* 2012). These studies assay thousands of variants simultaneously and provide detailed information on the relative mutational constraints on each base of a promoter, identifying transcription factor-binding sites and other regulatory elements.

Here, we use the methods of mutagenesis, selection, and sequencing to assay the fitness of nearly all single mutations in the *SUL1* promoter. We show that, while point mutations can increase fitness in sulfate limitation by up to ~10%, neither single mutations nor combinations of these mutations can increase fitness to the extent that *SUL1* amplification can. We also use these data to define potential transcription factor-binding sites that regulate *SUL1* expression and to identify point mutations that create new regulatory sites. Additionally, our assay is sensitive to post-transcriptional effects of *cis*-regulatory mutations and identifies detrimental mutations that create new upstream open reading frames in the *SUL1* 5' untranslated region.

Materials and Methods

Oligonucleotides, yeast strains, and plasmids used in this study

Oligonucleotides used in this study can be found in Supporting Material, Table S1. The *S. cerevisiae* strain used in this study was FY3, a *MATa* uracil auxotroph (*ura3-52*) of the S288c background. The strain deleted for *SUL1* was obtained by transformation with a PCR fragment containing a *NatMX* cassette and two flanking regions with homology to the *SUL1* locus. The transformant was backcrossed three times to FY2 to select for a clone containing both the *sul1* deletion and the *ura3-52* allele. A list of strains and plasmids used in this study can be found in Table S2.

Fitness estimates of individual strains

Fitness measurements of individual clones were performed as previously described (Payen *et al.* 2014) in sulfate-limited chemostats using a prototrophic FY strain where the *HO* locus had been replaced with *eGFP* (*MATa*: YMD1214).

Promoter truncation and mutagenesis

Promoter truncations were created by amplifying the *SUL1* locus from genomic DNA using oligos 314–319 as forward primers and oligo 266 as the reverse primer and the following PCR conditions: 98° for 1 min; then 25 cycles of 98° for 10 sec, 65° for 15 sec, and 72° for 15 sec; and then a final incubation at 72° for 5 min. These PCR products were cloned

into an *EcoRI*- and *SacI*-digested pRS416 vector by Gibson assembly, and their sequences were confirmed by Sanger sequencing. Yeast was transformed by lithium acetate transformation with each plasmid individually. At least two independent transformants were used for fitness measurements. The plasmid containing the 493-base promoter and remainder of *SUL1* was named “pMR002.”

The template used for mutagenesis was first amplified from pMR002 using oligos 295 and 297. Promoter mutagenesis was performed using the GeneMorph II Random Mutagenesis Kit (Agilent) according to manufacturer’s recommendations. One and 10 nanograms of template were amplified in mutagenesis reactions using oligos 295 and 297 and the following conditions: 95° for 2 min; then 25 cycles of 95° for 30 sec, 55° for 30 sec, and 72° for 30 sec; and then a final step of 72° for 10 min. These libraries were gel-extracted with a MinElute Gel Extraction kit (Qiagen), mixed one-to-one, and cloned into a *Bam*HI and *SacI*-digested pMR002 plasmid using Gibson assembly. One microliter of this reaction was transformed directly into ElectroMAX DH10B electrocompetent *Escherichia coli* (Life Technologies).

Barcodes were created by annealing oligos 283 and 296 and then performing a single cycle of extension using the Klenow fragment of DNA polymerase I (NEB). These fragments were cleaned and concentrated using a DNA Clean and Concentrator-5 kit (Zymo Research). pMR002 library plasmids were digested overnight with *Hind*III, dephosphorylated with calf intestinal phosphatase (NEB), and gel-purified. Barcode fragments were mixed in a 10-fold molar excess with cut library plasmids and cloned using Gibson assembly. ElectroMAX DH10B electrocompetent *E. coli* (Life Technologies) was transformed with 1 μ l of this reaction. In total, the library comprised 157,159 variants tagged by 634,639 barcodes. Yeast strains FY3 and FY3 *sul1* were transformed with barcoded plasmids using a high-efficiency protocol (Gietz and Schiestl 2007), resulting in >5 million transformed cells per library.

Continuous culture in chemostats

Nutrient-limited media (sulfate-limited and glucose-limited) were prepared as described (Gresham *et al.* 2008). The 200-ml chemostat vessels were inoculated with 1 ml of each pool ($\sim 2 \times 10^7$ cells). The pools were grown in chemostats for 25 hr in batch and then switched to continuous culture at a dilution rate of 0.17 ± 0.01 vol/hr at 30°. The cultures reached steady state after ~ 6 generations and were maintained for ~ 40 generations. A sample was taken immediately after the initiation of pumping and was designated generation 0 (G0). Samples (25–50 ml of culture at a density of 2×10^4 cells/ μ l) for cell counting and DNA extraction were passively collected once or twice daily, every three to six generations on average.

Mapping barcodes to promoter sequences

To map barcodes to promoter sequences, 25 ng of pMR002 was amplified for nine cycles with KAPA HiFi Hotstart Readymix using oligos 327 and 328 and the following cycling conditions:

98° for 20 sec and then nine cycles of 98° for 30 sec, 65° for 15 sec, and 72° for 25 sec. This reaction was cleaned with a DNA Clean and Concentrator kit (Zymo Research) and quantified using a Qubit fluorometer (Life Technologies). The products were sequenced with 300-base paired-end reads and a 12-base index read on a MiSeq (Illumina). Oligo 325 was used as the sequencing primer for both read 1 and read 2, and oligo 326 was used as the sequencing primer for read 3. Twelve-base barcode sequences were removed from read 1, and read 1 was trimmed to 270 bases using Prinseq (Schmieder and Edwards 2011). The resulting forward read and read 3 were merged and mapped to their consensus barcode (Hiatt *et al.* 2010; Starita *et al.* 2013). To remove unbarcoded plasmids and truncations occurring during Gibson assembly, barcode sequences were aligned to pMR002, and all barcodes that mapped were removed from the data set.

Barcode sequencing library preparation

Yeast samples (50 ml) were harvested and flash-frozen after growth in chemostats, and plasmids were extracted using the Zymoprep Yeast Miniprep II kit (Zymo Research). Sequencing libraries were created using two amplification steps. First, 4 μ l of each plasmid miniprep was amplified in a 25 μ l reaction by quantitative PCR with oligos 327 and 344 using KAPA HiFi HotStart ReadyMix (KAPA) on a Bio-Rad MiniOpticon (Bio-Rad). The following PCR conditions were used: 98° for 2 min and then cycles of 98° for 10 sec, 65° for 15 sec, and 72° for 25 sec. Reactions were stopped after 10–13 cycles to avoid overamplification. One microliter of each reaction was then diluted into a 25 μ l KAPA HiFi HotStart ReadyMix reaction with oligos P5 and NexV2ad2_N to add sequencing indices and amplify full-length products. As with the first reaction, quantitative PCR was used to avoid overamplification, and each reaction was stopped after either seven or eight cycles. The same cycling conditions were used as in the first amplification reaction. These reactions were cleaned with Ampure XP beads (Agencourt) and sequenced with 25-base reads on a MiSeq, NextSeq, and HiSeq 2000 (Illumina), using oligo 325 as the read 1 sequencing primer.

Barcode sequencing analysis

Twelve-base barcode sequences were extracted from reads and analyzed using custom software implemented in Python (File S1). Barcode sequences that perfectly matched the consensus barcodes were counted, and counts were converted to frequencies within each round of selection. Barcode frequencies were converted to log ratios between each round and the input. The fitness of each barcode was calculated as the slope of the ordinary least-squares regression of these ratios, and the number of generations elapsed for each sample. Variant fitnesses were normalized to wild-type fitness by subtracting the wild-type slope from the slope of each barcode. The fitness of a mutant is the average fitness of all barcodes that map to that mutant. To create a set of high-confidence variants, we compared the fitness and input read counts of the barcodes mapping to wild-type promoters. We qualitatively

set a minimum read count threshold at a point to maximize the number of wild-type barcodes yet minimize the variance of the wild-type fitness scores and dependence of fitness scores on input count number. Using this heuristic, we set the read count threshold for sulfate limitation experiments at 50 input reads and for glucose-limitation experiments at 15 input reads (Figure S1).

Creation and selection of a combinatorial library of high-fitness mutations

A library of promoters containing only combinations of five high-fitness mutations and neutral variation at one position was built using five sequential steps of PCR, each using the product of the previous reaction as a primer. Inadvertently, oligo 381 encoded either C or T, both neutral variants, at position -246. The primers and products used in each reaction were the following: (1) oligos 380 and 381; (2) the product from (1) and 237; (3) the product from (2) and 382; (4) oligos 237 and 382; (5) the product from (4) and M13F; (5) M13F and 237. All reactions were 25 μ l KAPA HiFi Hot-Start ReadyMix reactions using the following PCR conditions: 98° for 2 min; then 25 cycles of 98° for 10 sec, 65° for 15 sec, and 72° for 15 sec; and then 72° for 5 min. Only 15 cycles of PCR were used in reaction 5. The final product was gel-extracted using a MinElute Gel Extraction kit (Qiagen), cloned into pMR002 using Gibson assembly, used to transform ElectroMAX DH10B electrocompetent *E. coli* (Life Technologies), and then used to transform YMD3017 using a high-efficiency yeast transformation (Gietz and Schiestl 2007). Selections and plasmid extractions were performed as before. Fragments for sequencing were amplified using 400 nM each of a custom forward index primer (oligos 395–402, one per sample) and oligo 328, using the following PCR conditions: 98° for 2 min and then cycles of 98° for 10 sec, 65° for 15 sec, and 72° for 25 sec. Quantitative PCR was used to avoid overamplification, and reactions were stopped after 8–16 cycles, depending on the sample.

Libraries were sequenced with overlapping 150-base reads on a Miseq (Illumina) using oligo 442 as a forward primer and oligo 387 as a reverse primer. These reads covered five of the six mutated sites in the library. We sequenced the final mutated site with an 8-base index read (using oligo 443) and used the second index read to demultiplex samples. Reads were genotyped at each mutated site and analyzed similarly to our barcode sequencing data (code for this analysis can be found in File S1).

Generation and genomic integration of promoter variants

We first replaced the genomic *SUL1* promoter with *URA3*. *URA3* was amplified from FY3 genomic DNA with oligos 678–681, and this PCR product was used to transform YMD3107 via high-efficiency yeast transformation (Gietz and Schiestl 2007). We created single mutants of the *SUL1* promoter in two sequential PCR steps. We amplified the 3' end of the promoter with a primer containing each mutation (oligos 517–522 and 603–605) and oligo 237, using pMR002

as a template. These products were gel-extracted with a Fast-Gene PCR Extraction Kit (Nippon Genetics) and used as primers along with M13F to add the 5' end of the promoter to the fragment (using pMR002 as template). These reactions were gel-extracted and cloned into pMR002 by Gibson assembly. The promoter-*SUL1* fragments of these plasmids were amplified with oligos 237 and 685 and used to transform YMR002, creating strains YMR008-YMR012 and YMR017-YMR020. These strains were made prototrophic (URA⁺) by backcrossing to YMD3018, creating strains YMR022 and YMR024-YMR030.

Matching fitness data to known transcription factor motifs

Single-mutant log-normalized frequencies at 0 and 40 generations were estimated using the slope and y-intercept of the linear fit used to calculate variant fitness, creating a log-likelihood of finding each variant in the data set. These ratios were then ordered by position and identity (*i.e.*, A, C, T, or G). The estimated wild-type ratio at 40 generations was used for the wild-type base at each position, and a value of 1.0 was used for missing data. Any extrapolated values <0.001 were set to 0.001. Specific position ranges were extracted from this matrix and compared to known transcription factor motifs (Zhu and Zhang 1999; MacIsaac *et al.* 2006) using Tomtom version 4.10.0 (Gupta *et al.* 2007).

Finding newly created transcription factor-binding sites

We enumerated all possible single mutations in the *SUL1* promoter in the context of a 25-mer (each mutation was flanked by 12 upstream and downstream wild-type bases), as well as their wild-type alternatives. We searched this set of sequences for occurrences of motifs with Find Individual Motif Occurrences (FIMO) (Grant *et al.* 2011), relaxing the significance threshold for reporting motif matches to $P = 0.01$. This relaxed threshold allowed us to identify even weak matches to a motif that could be improved by point mutations. We compared the significance of each mutant match to the significance of that motif's match to the wild-type 25-mer for all motif matches that overlapped the middle position in the sequence and calculated a log-normalized ratio of these two scores.

Data availability

Scripts can be found at <https://github.com/msr2009/Rich2016>. Raw data can be found in File S2. Raw sequencing reads can be found in the National Center for Biotechnology Information Sequence Reads Archive, Bioproject ID PRJNA273419. A flow-chart describing the experiments presented here can be found in Figure S7. Variants are named in this manuscript using HGVS format. For example, the mutation at position -458 from T to A is named -458T>A.

Results

Extent of the *SUL1* promoter

We first sought to define the *SUL1* regulatory region because little functional annotation is available apart from the identification of three putative TATA boxes at -199, -93, and

–91 (Kellis *et al.* 2003; Basehoar *et al.* 2004). We created a centromeric plasmid-borne copy of *SUL1* that included the coding region, 687 bases upstream of the start codon and 270 bases downstream of the termination codon. This region (chromosome II: 788548–792107) begins 118 bases downstream of the *VBA2* termination codon and stops 57 bases upstream of the first base of *ARS228*. Deletion of this genomic region causes a 19% decrease in sulfate-limited fitness (Payen *et al.* 2015). Expression of plasmid-borne *SUL1* in the deletion background increased fitness by 20.4%. To identify the functional extent of the *SUL1* promoter, we measured the fitness in sulfate-limited media of *sul1Δ* yeast strains transformed with plasmids containing one of a set of truncated promoters, each progressively deleting ~100 bases (Figure 1). Truncation to a 592-base promoter caused a small fitness decrease (–6.45%) compared to the full-length regulatory sequence, although truncation further to 493 bases restored fitness. All truncations that created a promoter shorter than 493 bases were unable to complement the deletion of the endogenous *SUL1*, leading to substantial fitness defects (–33% to –53%). We defined the 493 bases (chromosome II: 788742–789234) as the minimal *SUL1* promoter and used this region for mutagenesis. Since the transcription start site of *SUL1* is also unannotated, mutations will be numbered relative to the *SUL1* start codon; for example, –493 is chromosome II: 788742 and –1 is chromosome II: 789234.

SUL1 promoter mutagenesis and selection

We used error-prone PCR to create random mutations in the *SUL1* promoter and cloned these promoter variants scarlessly by Gibson assembly into the *SUL1* construct. After uniquely barcoding plasmids and linking the plasmid barcodes to the sequence of the promoter variants, we obtained sequences of 152,723 variants uniquely tagged by 630,517 barcodes. Each uniquely barcoded variant had on average 2.2 mutations. The wild-type promoter was tagged by 92,753 barcodes (15.1% of the total barcodes) (Figure S2A). We transformed both *SUL*⁺ and *sul1Δ* strains with this library. Transformants were competed as a pool during growth in sulfate-limited chemostats, and the activity of each barcoded promoter, as approximated by the relative fitness of each strain in the pool, was calculated as described in *Materials and Methods*. After stringent filtering for high-confidence variants (*Materials and Methods*), we calculated a wild-type-normalized fitness value for 29,906 promoter variants. The pooled transformants as a whole had a median fitness decrease of –1.7%. Single-mutant fitnesses were specific to sulfate limitation, as fitness values after selection in glucose limitation, in which *SUL1* activity does not drive competitive fitness, were neutral (Figure S4), and the distribution of all fitness scores in glucose limitation was not statistically different from the fitness score of the 94,912 wild-type promoter sequences in sulfate-limited medium ($P = 0.38$, t -test).

Both the distribution of wild-type barcodes and the distribution of all variants were centered on zero and had a long negative tail (Figure S2B). Unlike the distribution for wild-type

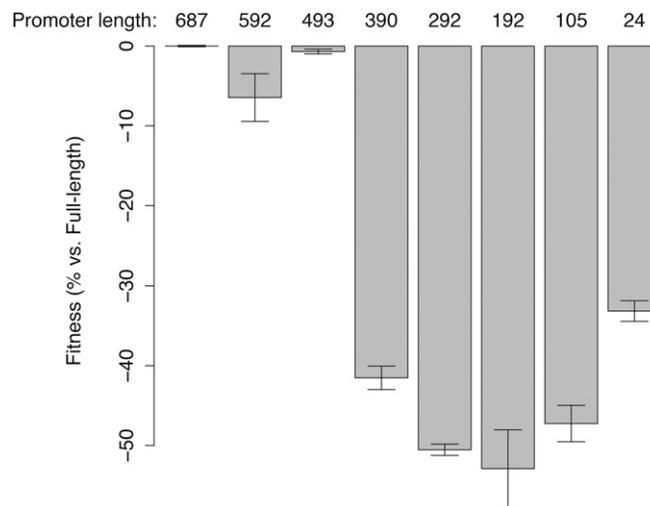


Figure 1 Determining the extent of the *SUL1* promoter by truncation. The fitness of *sul1Δ* strains harboring plasmids with full-length and truncated *SUL1* promoters was assayed by competition in sulfate-limited chemostats against a wild-type (*SUL*⁺) strain. Fitness values are normalized to the full-length promoter (687 bases). Promoter lengths are measured from the *SUL1* start codon. Error bars represent standard deviation ($n = 2$).

barcodes, the distribution of variant fitness had a large shoulder corresponding to variants with fitness decreases between –15 and –5%, consistent with many positions in the promoter being sensitive to mutation.

We also examined the effect of mutations in a strain in which the endogenous copy of *SUL1* was not deleted. The fitness effects of variants were generally highly correlated between this *SUL*⁺ strain and the *sul1Δ* strain (Spearman's $\rho = 0.859$) (Figure S3). Fitness values were correlated between the two backgrounds for variants with wild-type-normalized fitness > –15%. However, variants with wild-type-normalized fitness < –15% in *sul1Δ* were not as unfit in a *SUL*⁺ background (Figure S3).

Effect of point mutations in the *SUL1* promoter

SUL1 amplification is present in all populations at generation 100 during laboratory experimental evolution in sulfate limitation (Gresham *et al.* 2008; Miller *et al.* 2013; Payen *et al.* 2014). Accumulating more than one mutation in the 493 bases of the *SUL1* promoter during those 100 generations is unlikely, so we first limited the bulk of our analysis to the fitness effects from single mutations. We assayed 1400 of the 1479 (94.6%) possible single mutants of the *SUL1* promoter (Figure 2). Most mutations had little effect on fitness. Based on the fitness distribution of uniquely barcoded wild-type promoters, we established an empirical false discovery rate of 5% to be wild-type normalized fitness less than –7.1% and greater than 4.3%. Only 11 single mutations increased fitness >4.3%, whereas 50 single mutations decreased fitness by >7.1%. Overall, single mutations had a narrower distribution than the set of all variants, including a much shorter negative tail. The effect size of fitness-increasing mutations was similar between singly and multiply mutated variants;

only 46 variants with multiple mutations had fitness increases greater than the maximum single-mutant fitness (9.4%), and these variants increased fitness up to 11.7%. All 46 of these variants contained at least one of $-353\text{T}>\text{G}$, $-372\text{T}>\text{C}$, $-404\text{C}>\text{T}$, and $-458\text{T}>\text{A}$, and 39 contained $-372\text{T}>\text{C}$, the second-most-fit variant in the library. The effects of fitness-decreasing variants appeared to be additive, as 1054 multiply mutated variants decreased fitness more than the most detrimental single mutation (-14.7%) (Figure S2B).

Of the 50 single mutations that decreased fitness by $>7.1\%$, 46 were found in four sites in the promoter (A–D in Figure 2B). These sites appear as low-fitness “stripes” on the heatmap, with nearly every mutation that occurs in each site decreasing fitness. These sites are short, between 8 and 20 bases, and likely correspond to binding sites for the transcription factors that regulate *SUL1* expression (see below). Truncation of the *SUL1* promoter to 393 bases, which removes two of these sites, resulted in the failure of the shortened promoter to complement a *SUL1* deletion, presumably because it provides insufficient expression.

Discovering *SUL1* transcriptional regulatory sites

We used the single-mutation data to identify the likely transcription factors that regulate *SUL1* expression. The wild-type sequence of site D (-200 to -193) is TATAAATA, matching a canonical yeast TATA box (Basehoar *et al.* 2004). Using the slope and intercept data for each single mutant in the site, we created a matrix of the log-likelihoods of finding each single mutant in the site after 40 generations. We searched transcription factor motif data sets (Zhu and Zhang 1999; MacIsaac *et al.* 2006) to find significant matches to the log-likelihood matrix. This analysis identified *Spt15*, the yeast TATA-binding protein, as a match for this site (Figure 3D), albeit one with a high false-discovery rate ($q = 0.27$), likely driven by missing data in the first three positions of the site. Because we conservatively assigned missing values a log-likelihood of 1, *i.e.*, no change in frequency during the selection, we decreased the significance of our matches to this site.

We then applied this methodology to the other three sensitive sites (Figure 3, A–C). The log-likelihood matrix for site A, spanning positions -465 to -448 , weakly matched the motif for *Cbf1*, a known regulator of sulfate metabolism that recognizes CACGTG (Dowell *et al.* 1992), and *Tye7*, a glycolytic activator that is implicated in Ty1-mediated gene expression. A second site, from positions -451 to -449 , flanking a hypothetical *Cbf1* motif (CTCGTG), was also sensitive to mutation. These positions partially match the RYAAT motif, which is necessary for full induction of genes during growth in limited sulfate by enhancing binding of the *Cbf1*–*Met28*–*Met4* regulatory complex (Siggers *et al.* 2011).

The log-likelihood matrices for both site B and site C contain the CCACA motif recognized by *Met31* and *Met32* (Blaiseau *et al.* 1997). Site B matched *Met32* ($q = 0.03$) and *Met31* ($q = 0.09$) and appears to be an incomplete palindrome [GCCACA(CG)TGTGGC] centered on position -407 . Site C

was the most significant match of the analysis and matched the *Met32* motif ($q = 0.005$).

Although sites A, B, and C were generally sensitive to most mutations, some mutations yielded large ($>5\%$) increases in fitness. Comparison of the wild-type sequence to the highest-fitness variant at these positions showed that each of these binding sites was one mutation away from the consensus binding site for the implicated transcription factor. For example, mutation $-458\text{T}>\text{A}$ creates the consensus binding site of *Cbf1* (CACGTG) and increased fitness by 6.3% in our competition and increased *Cbf1* binding strength ~ 140 -fold, as measured *in vitro* by the MITOMI assay (Maerkl and Quake 2007). Mutations $-413\text{A}>\text{G}$ and $-404\text{C}>\text{T}$ had similar effects in site B, creating a consensus binding site for *Met32* and increasing fitness 6.5 and 7.1%, respectively. Mutation $-353\text{T}>\text{G}$, which also creates a consensus *Met32*-binding site, caused the largest fitness increase in the data set (9.4%). The importance of $-413\text{A}>\text{G}$ and $-356\text{T}>\text{G}$ distinguishes these sites as *Met32*-binding sites, rather than similar *Met31*-binding sites, as the importance of a 5' guanine is found only in the *Met32* site.

Other short sites in the promoter (*e.g.*, -329 to -321) showed similar trends of 8–20 contiguous positions being sensitive to mutation, but with smaller fitness effects. When we performed a search for the log-likelihood of these sites against known transcription factor-binding sites, we found no significant matches (data not shown).

Creation of new regulator-binding sites

The methodology that we used to find endogenous regulators depends on a signature of purifying selection throughout a binding site and so is not generally applicable to finding mutations that create new transcription factor-binding sites. Endogenous *SUL1* regulatory sites are highly sensitive to mutation across the entire site, except for rare mutations that optimize active sites (sites A, B, and C in Figure 2B). This signature of purifying selection would not be found in inactive binding sites, and point mutations that activate these sites would be independent of the surrounding positions. As such, we took a different approach to find binding sites that occur or are lost upon single mutation.

To identify such sites, we first enumerated all 25-mers and all possible single mutations centered in 25-mers *in silico* and then searched these 25-mers for matches to transcription factor-binding sites using relaxed parameters to allow for weak matches. In total, 124,316 motifs matched at least one 25-mer at a threshold of $P < 0.01$. An arbitrary P -value of 0.01 was used for motifs that matched either the mutant or wild-type 25-mer, but not both. We calculated the log ratio between P -values for the wild-type and a mutant motif match, $\log_2(P_{\text{wt}}/P_{\text{mut}})$ and used this ratio as a measure of motif strength.

Of the 588 motifs with a match significance of $P < 0.0001$, 347 were either strengthened or weakened at least 10-fold by mutations (Figure S5). The majority (52/56) of mutations that decreased fitness by $>5\%$ and altered the significance

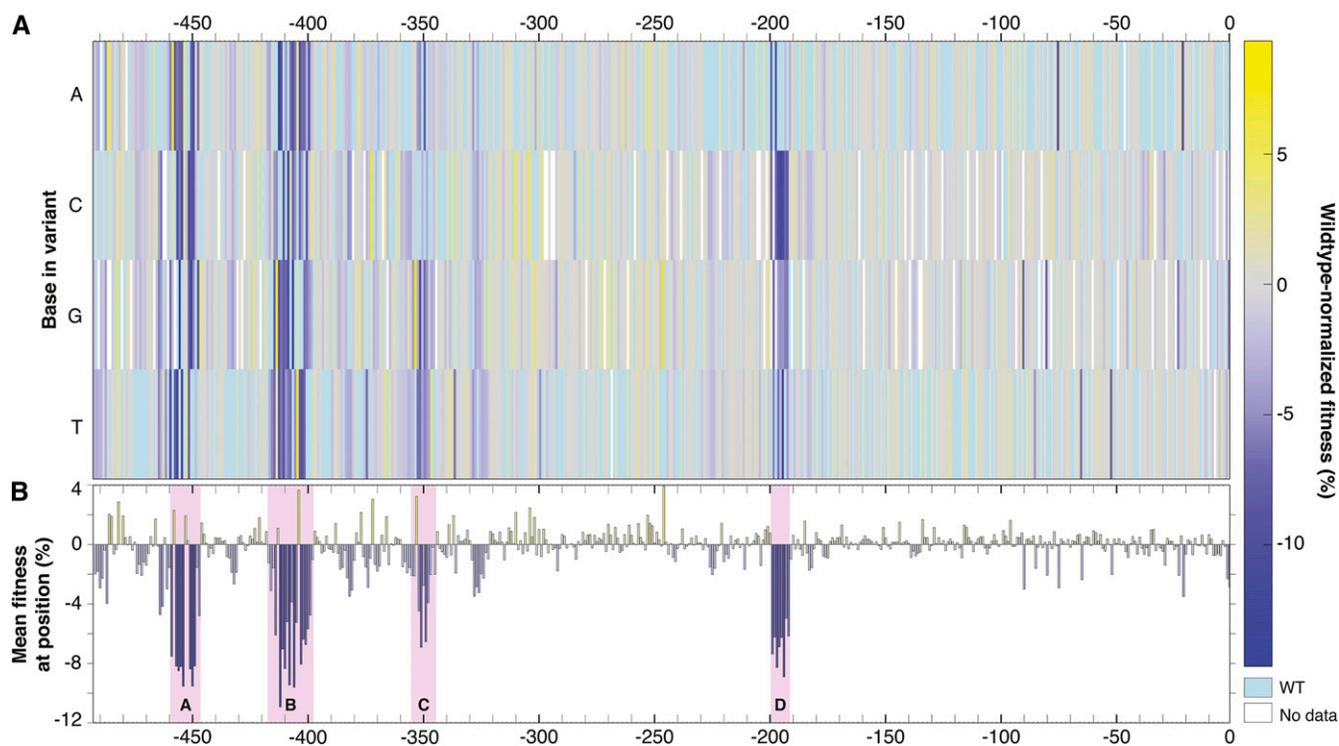


Figure 2 The effect of single mutations in the *SUL1* promoter on sulfate-limited fitness. (A) The fitness change of each single mutation in our data set is plotted as a heatmap. At each position in the promoter along the x-axis, the base present in each single mutant is ordered on the y-axis, and every cell is shaded based on the fitness change of that mutation. Missing data are white cells in the heatmap, and the wild-type base at each position is teal. (B) Mean fitness change of all variants that are mutated at that position. Sites examined in Figure 3 are marked with pink backgrounds.

of a motif match by at least 10-fold occurred in either the TATA box or sites A–C or created a new upstream open reading frame in the 5' UTR. Seven mutations matching 11 different motifs increased fitness by >5% and had at least a 10-fold increase in motif significance when compared to wild type. We identified three of these (–353T>G, –404C>T, and –458T>A) in sites C, B, and A, respectively, as mutations that optimize endogenous *Met32*- and *Cbf1*-binding sites. One of the remaining mutations, –372T>C, creates a *Met32*-binding site (GCCACA), increasing the significance of the motif match by 23-fold and fitness by 9.3%, the second-highest fitness increase from a single mutation in our data set. Another, –246T>G, creates a one-off *Cbf1* site (CATGTG), increasing the significance of the motif match by 32-fold and increasing fitness by 8.4% (Figure S5B).

Other mutations that increased fitness >5% do not have clear-cut biological explanations. For example, –310A>T increased fitness by 5.5% and the significance of the match to *Hap2*, a glycolytic activator, by 20-fold. The –482A>G mutation might strengthen the binding of *Aft1* or *Aft2* (22.5- and 96-fold, respectively) or weaken the binding of *Xbp1* by 43-fold. *Aft1* and *Aft2* regulate iron homeostasis, and *Aft1* interacts physically with *Cbf1* (Measday *et al.* 2005); the fitness increase caused by –482A>G may be dependent on *Cbf1* binding to the neighboring binding site (site A). Alternatively, the fitness increase could also be caused by decreased *Xbp1*-mediated repression, as *Xbp1* is known to repress

other *Met4*-responsive genes (Mai and Breeden 1997; Lee *et al.* 2010).

Effect of upstream open reading frames on *SUL1* fitness

Short open reading frames starting upstream of a gene's coding sequence can post-transcriptionally regulate gene expression, as the ribosome creates unproductive polypeptides instead of the correct protein (Morris and Geballe 2000; Yun *et al.* 2012). No upstream open reading frames are present in the wild-type *SUL1* 5' untranslated region. Because our assay selects on the amount of *Sul1* protein in the cell, it should identify mutations that cause a fitness defect due to post-transcriptional regulation of *SUL1* expression. Therefore, we searched for mutations that create an upstream open reading frame (Figure 4), *i.e.*, the mutation of a 3-mer to an AUG start codon within the 5' untranslated region of the *SUL1* transcript (Smith *et al.* 1995). Our data contained 23 of the 26 possible upstream ORFs, 8 of which decreased fitness by at least 5%; none of the upstream ORF mutations increased fitness. Four upstream ORFs created a polypeptide that was out of frame with the *SUL1* coding sequence and read past the *SUL1* start codon. These mutations were invariably deleterious (decreasing fitness by $\geq 6\%$). Six upstream ORFs created an in-frame fusion to *Sul1*, with lengths ranging from 2 to 18 amino acids. The longest fusion decreased fitness by 7.2%, and as the fusions shortened, their effect on fitness decreased, with fusions adding

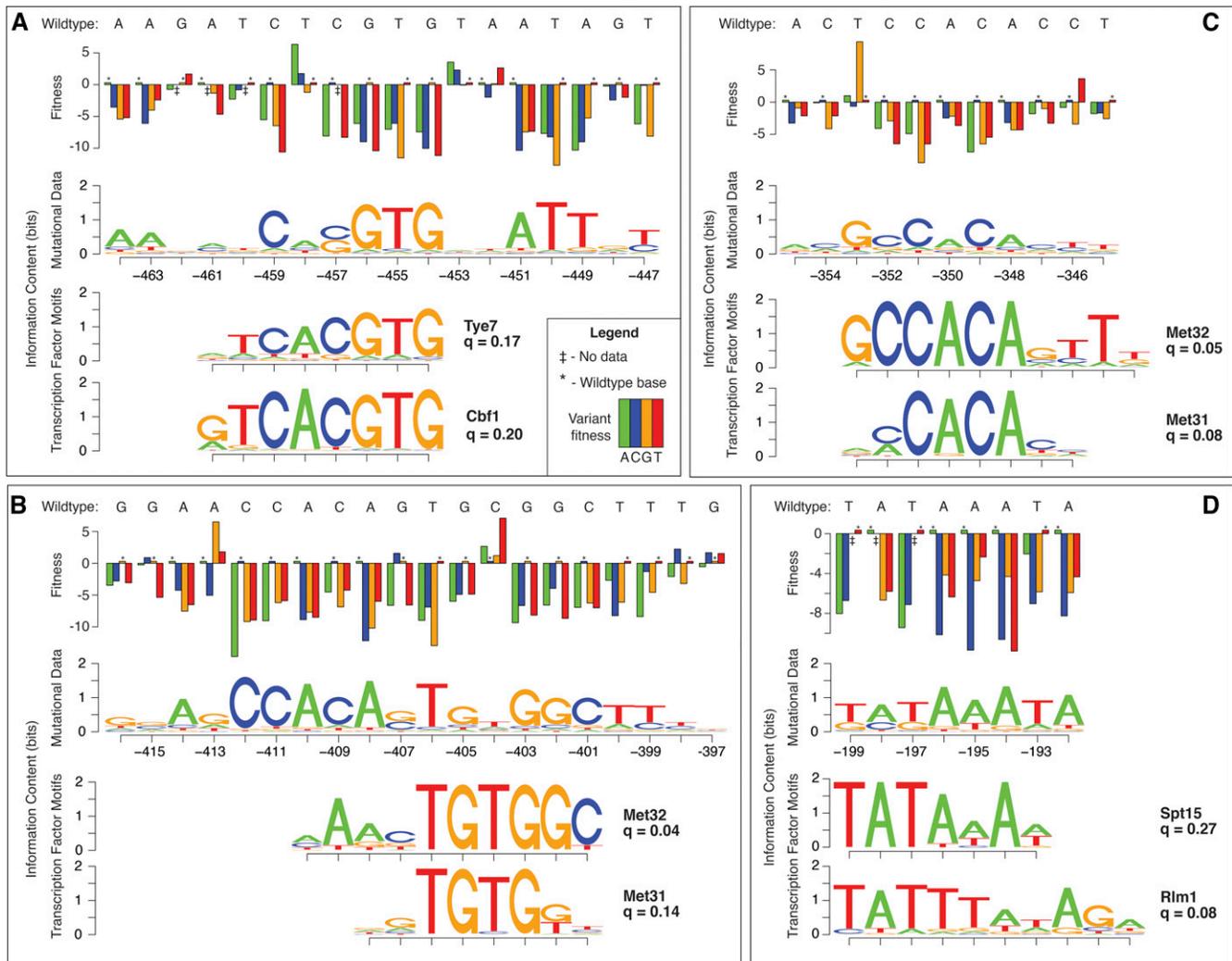


Figure 3 Determining *SUL1* transcriptional regulatory sites through mutagenesis. To identify the transcription factors binding at the four sites highlighted in Figure 2B, we compared position weight matrices based on the fitness values for single mutations at each site with a database of known yeast transcription factor-binding motifs. (A–D) The wild-type sequence for each region is shown at the top of each panel. Below this is a barplot showing the fitness value for the single mutations at each site. Below the barplot is a set of sequence logos, the first of which is based on the mutational data for the region (calculated as described in *Materials and Methods*). Underneath this logo are the significant ($q < 0.2$, except in D) matches between the fitness-based PWM and yeast transcription factor-binding sites. Panel labels are consistent with the labels in Figure 2B.

five or two amino acids being neutral (0.3 and 1.3% fitness increases, respectively).

Combinatorial analysis of high-fitness mutations

Six mutations in our data set yielded a $>6\%$ increase in fitness: $-353\text{T}>\text{G}$, $-372\text{T}>\text{C}$, $-246\text{T}>\text{G}$, $-404\text{C}>\text{T}$, $-413\text{A}>\text{G}$, and $-458\text{T}>\text{A}$. Based on our analyses above, these mutations may increase the affinity of a transcriptional activator. Individually, these mutations conferred $<10\%$ fitness increases, much less than the increased fitness ($>37\%$) of evolved strains with amplifications containing *SUL1* (Gresham *et al.* 2008; Payen *et al.* 2014). No variants that passed filtration combined two or more of these mutations. To investigate the combinatorics of these mutations, we created another library in which five of these sites (excluding $-246\text{T}>\text{G}$) were present as either the reference base or

the high-fitness mutation. Our library contained all 32 possible mutation combinations. As before, we transformed a *sul1Δ* strain with this library, competed the population in sulfate-limited chemostats, and calculated variant fitness.

Fitness effects from single mutations were additive when combined in pairs, although double mutants increased fitness only 10.5%, a fitness increase of 1.15% above the most-fit single mutant (Figure 5). The addition of more mutations to each variant did not increase fitness significantly above the fitness of the double mutants, with 11% the maximum increase in fitness reached by any mutation combination.

Either $-353\text{T}>\text{G}$ or $-372\text{T}>\text{C}$ were necessary to reach the fitness plateau of an $\sim 10\%$ increase. Four combinations without these mutations had a wild-type-normalized fitness increase of $8.4 \pm 0.2\%$ on average. All combinations

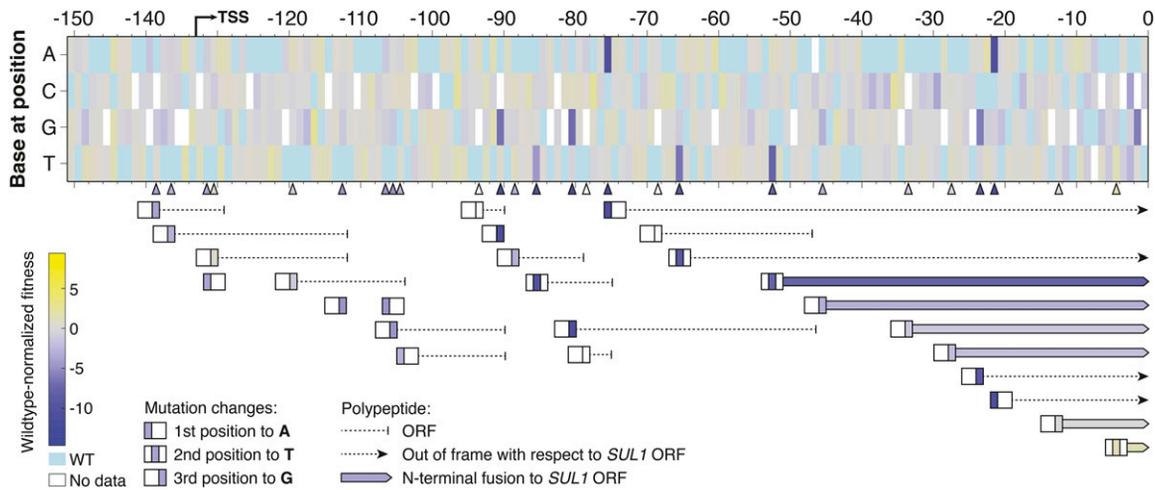


Figure 4 The effect of upstream open reading frames on *SUL1* expression. We searched the 150 bases upstream of the *SUL1* start codon for sites that could mutate to form an upstream ATG. The heatmap of this region is shown with triangles marking the positions of mutations to upstream start codons. Below each triangle is a cartoon representation of the upstream open reading frame created by each ATG. The position of the colored box in each cartoon represents which base was mutated to create an ATG, and each box is colored based on the fitness change of that mutation. Dashed lines extending to right of each box show the extent of the hypothetical polypeptide encoded by each ORF. In the case in which the ORF is in-frame with the *SUL1*-coding sequence, these polypeptides are drawn as colored boxes.

including $-353\text{T}>\text{G}$ or $-372\text{T}>\text{C}$ ($n = 26$) had a wildtype-normalized fitness increase of $10.4 \pm 0.3\%$.

We tested a set of eight single mutants and the combinatorial variant in their native genomic context. Integration of the 493-base promoter that deletes ~ 200 bases of the intergenic sequence *SUL1* and *VBA2* led to strain fitness of approximately that of a *sul1* Δ strain (data not shown). We therefore integrated eight mutations and the combinatorial allele in the context of the entire 687-base promoter. The integrations resulted in a scarless replacement of the *SUL1* promoter. The fitnesses of integrated single mutants were highly correlated with pooled measurements of fitness (Pearson coefficient = 0.872, $P = 0.002$, Figure S6A). All variants except $-246\text{T}>\text{G}$ were contained within a 95% confidence interval based on the pooled fitness values of the barcodes mapping to each variant (Figure S6B). We could not calculate a confidence interval for the five-mutation variant because it was not found in the barcoded library and the combinatorial library did not employ barcodes.

Discussion

Amplification has been proposed to be a “quick and dirty” solution for adaptation, allowing more stable mutations to be selected over a longer period of time (Hendrickson *et al.* 2002; Yona *et al.* 2012); indeed, amplification of stress-specific genes may be detrimental during nonstress growth. The fitness increase due to *SUL1* amplification [up to 51% (Payen *et al.* 2014)] is much higher than that due to any of the single mutations, the largest increase in fitness of which was 9.35%. Single mutations were not additive when combined, and the fitness increase of any combination of high-fitness mutations plateaued at $\sim 11\%$. This lack of additivity may imply that there is a limit to the maximal rate of transcription achievable

from the *SUL1* promoter, limiting the effect of *cis*-regulatory mutations. Thus, for the *SUL1* promoter, amplification and *cis*-regulatory mutations are not equivalent, which is inconsistent with the “quick and dirty” hypothesis of amplification. Two coding mutations in *SUL1* have been isolated that increase sulfate-limited fitness by 23%, although it is not known how they affect protein function. These mutants were isolated in a strain in which gene amplification was prevented by deletion of the recombinase gene *RAD51* (Payen *et al.* 2015). It remains to be tested whether beneficial mutations in the 3' UTR can surpass gene amplification.

In this study, we did not investigate the potential for *trans*-acting mutations to increase *SUL1* expression, in part because the transcription factors governing *SUL1* expression were largely unknown. Our data show that there are three sites in the *SUL1* promoter that match binding sites for sulfate-regulatory transcription factors, two for *Met32* and one for *Cbf1*. Both these transcription factors provide DNA-binding specificity for *Met4*, which recruits other transcriptional machinery. *Met4* is a strong transcriptional activator (Titz *et al.* 2006), such that little *Met4* occupancy in the promoter, perhaps only a single binding event, may be sufficient to near-maximally activate *SUL1*. Mutations that create new binding sites for *Met32* or *Cbf1* presumably add an additional site for *Met4* occupancy and conferred fitness increases up to 9.3%, which is a much smaller effect than that of amplification. Amplification would increase the total copies of *SUL1* available to be transcribed, leading to more protein than from a single copy.

We determined that the *Met32*- and *Cbf1*-binding sites are the primary regulators of *SUL1* expression and that these sites are not the consensus binding site for either factor. Consistent with our analysis, *met32* strains have decreased fitness under sulfate limitation, whereas *met31* strains do not (Payen *et al.*

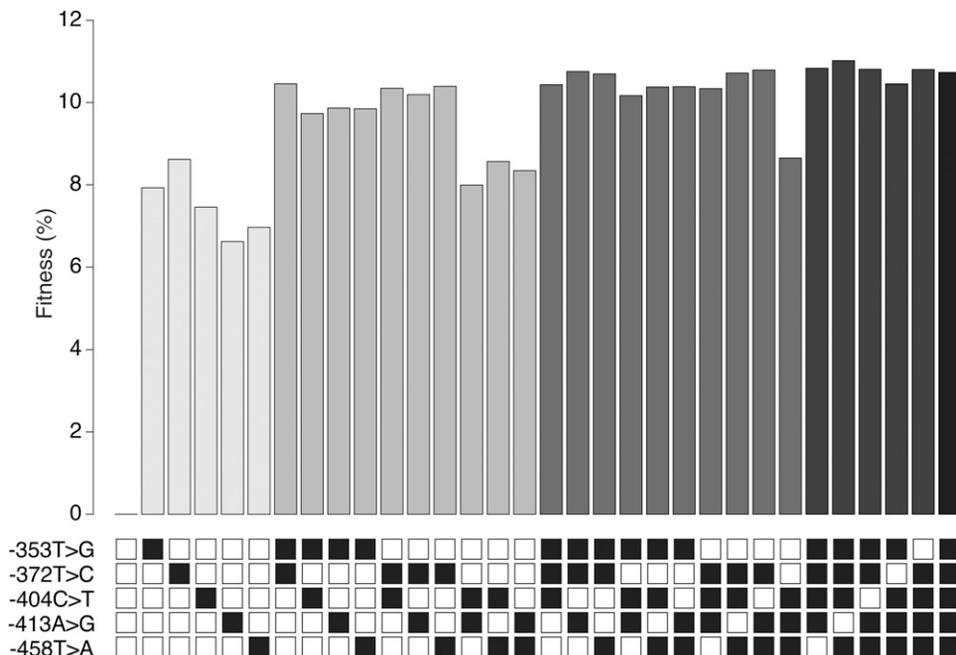


Figure 5 Combinatorial effects of high-fitness mutations. The wild-type-normalized fitness for each of 32 combinations of five high-fitness mutations is shown as a bar. The genotype of each variant is identified on the x-axis: a black box in each row indicates the presence of each mutation in that variant.

2015). Our data also strengthen the observation that *Met31* and *Met32* are not fully redundant (Su *et al.* 2008; Cormier *et al.* 2010). Because each binding site in the *SUL1* promoter is at least one mismatch away from its consensus site, their annotation by sequence-based motif-finding methods is difficult. Studies showing combinatorial control by *Cbf1* and *Met31/32* of *Met4*-dependent genes failed to identify by standard motif-finding approaches the *Cbf1*-binding site that we identified in the *SUL1* promoter (Lee *et al.* 2010; Carrillo *et al.* 2012), although combinatorial regulation of *SUL1* can be parsed by analyzing transcription after inducing or repressing sulfate-metabolism regulators (McIsaac *et al.* 2012; Petti *et al.* 2012). We searched the *SUL1* promoter sequence using two yeast transcription factor databases, YEASTRACT (Teixeira *et al.* 2006; Monteiro *et al.* 2008; Abdulrehman *et al.* 2011; Teixeira *et al.* 2014) and YeTFaSCo (de Boer and Hughes 2012), to compare our results to the results from *in silico* analyses. Our search yielded 99 possible motifs, only one of which, a *Met31* motif centered at -410 (overlapping a site that we hypothesize to be a *Met32* site), aligned with any of our predictions.

Why are none of the transcription factor-binding sites in the *SUL1* promoter consensus sequences? A 5% increase in fitness during sulfate limitation would be a strong evolutionary force, and these mutations should therefore become fixed in a population. While the effects of these mutations were specific to sulfate limitation, they do not appear to be detrimental under permissive, sulfate-rich conditions, and thus they would not be under purifying selection during nonstressful growth. However, sulfate limitation, at least to the extent of the experimental design here, may not be a selective pressure frequently experienced by yeast in the wild. Alternatively, balancing selection may be working against maximal activation of *SUL1*; as another activity, *Sul1* also transports toxic

heavy metals (Cherest *et al.* 1997). *SUL1* also appears to be dispensable under rich media conditions, as lager strains carry loss-of-function mutations in the gene (Libkind *et al.* 2011). Our assay measures fitness in steady state, so it is possible that the wild-type promoter may have favorable kinetic qualities for fast adaptation to sulfate-limited environments; any temporally dependent aspects of *SUL1* regulation would go uninvestigated by our methodology.

The use of plasmids may confound some of our results. A 493-base *SUL1* promoter on a centromeric plasmid complemented the genomic deletion of *SUL1*, but not when it was genomically integrated. Centromeric plasmids have a low copy number, but not necessarily a copy number of 1 (Karim *et al.* 2013). An increase in plasmid and *SUL1* copy number may have masked some of the effects of truncation and mutagenesis of the *SUL1* promoter, with multiple copies of low-fitness promoters complementing a *SUL1* deletion. We saw this effect when measuring the fitness of variants in the context of a *SUL⁺* strain, where the endogenous copy of *SUL1* masked variants that were detrimental in a *sul1Δ* background, and this could be the cause of the discrepancy in the plasmid and integrated fitness measurements for variants like $-348A>T$. Promoter strength has been shown to affect plasmid copy number (Karim *et al.* 2013). Our assay essentially counts the number of plasmids in a population, so biasing plasmid copy number dependent on promoter strength would also bias our results.

Alternatively, background mutations in the strain may alter fitness independently of the *SUL1* promoter genotype. Transformation is known to be mutagenic (Shortle *et al.* 1984), and mutations can accumulate during strain construction (Wilkening *et al.* 2014). We advise that future studies carefully control for plasmid copy number. Copy number is not an issue with integrations, but minor fitness effects of variants

may be difficult to discern in integrants above the effects of extraneous mutations in the strain background. This problem may be circumvented by obtaining large numbers of bar-coded integrants.

By using chemostats, we were able to perform sensitive analyses of the fitness effects of *cis*-regulatory mutations, an analysis not accessible to other methodologies that assay only transcriptional output. This approach should be applicable to studying transcriptional regulation more generally, as the *SUL1* promoter can be replaced with other *S. cerevisiae* promoters, allowing for the same mutational analysis as long as these other promoters are functional during growth under limited sulfate. Finally, because this approach is sensitive to post-transcriptional effects on protein expression, it could be adapted as a platform for studying post-transcriptional regulation in yeast, allowing, for example, assays of the effects of 5' untranslated regions on protein expression.

Acknowledgments

We thank Bill Noble and Charles Grant for help with Tomtom and FIMO and for discussions about motif matching and Kunihiro Ohta for the use of his tetrad dissection microscope. This work was supported by grants R01 GM094306 (to M.J.D.) and P41 GM103533 (to M.J.D. and S.F.) from the National Institute of General Medical Sciences of the National Institutes of Health. M.J.D. is a Rita Allen Foundation Scholar and a Senior Fellow in the Genetic Networks program at the Canadian Institute for Advanced Research. S.F. is an investigator of the Howard Hughes Medical Institute. M.S.R. was partially supported by a Short-term Postdoctoral Fellowship from the Japan Society for the Promotion of Science.

Literature Cited

Abdulrehman, D., P. T. Monteiro, M. C. Teixeira, N. P. Mira, and A. B. Lourenço *et al.* 2011 YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic Acids Res.* 39: D136–D140.

Basehoar, A. D., S. J. Zanton, and B. F. Pugh, 2004 Identification and distinct regulation of yeast TATA box-containing genes. *Cell* 116: 699–709.

Blaiseau, P. L., A. D. Isnard, Y. Surdin-Kerjan, and D. Thomas, 1997 Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol. Cell. Biol.* 17: 3640–3648.

Brison, O., 1993 Gene amplification and tumor progression. *Biochim. Biophys. Acta* 1155: 25–41.

Carrillo, E., G. Ben-Ari, J. Wildenhain, M. Tyers, D. Grammentz *et al.*, 2012 Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Mol. Biol. Cell* 23: 1928–1942.

Cherest, H., J. C. Davidian, D. Thomas, V. Benes, W. Ansorge *et al.*, 1997 Molecular characterization of two high affinity sulfate transporters in *Saccharomyces cerevisiae*. *Genetics* 145: 627–635.

Cormier, L., R. Barbey, and L. Kuras, 2010 Transcriptional plasticity through differential assembly of a multiprotein activation complex. *Nucleic Acids Res.* 38: 4998–5014.

de Boer, C. G., and T. R. Hughes, 2012 YeTFaSCO: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.* 40: D169–D179.

Dowell, S. J., J. S. Tsang, and J. Mellor, 1992 The centromere and promoter factor 1 of yeast contains a dimerisation domain located carboxy-terminal to the bHLH domain. *Nucleic Acids Res.* 20: 4229–4236.

Gietz, R. D., and R. H. Schiestl, 2007 High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* 2: 31–34.

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

Gresham, D., M. M. Desai, C. M. Tucker, H. T. Jenq, D. A. Pai *et al.*, 2008 The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* 4: e1000303.

Gruber, J. D., K. Vogel, G. Kalay, and P. J. Wittkopp, 2012 Contrasting properties of gene-specific regulatory, coding, and copy number mutations in *Saccharomyces cerevisiae*: frequency, effects, and dominance. *PLoS Genet.* 8: e1002497.

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

Hendrickson, H., E. S. Slechta, U. Bergthorsson, D. I. Andersson, and J. R. Roth, 2002 Amplification-mutagenesis: evidence that “directed” adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proc. Natl. Acad. Sci. USA* 99: 2164–2169.

Hesselberth, J. R., X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom *et al.*, 2009 Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* 6: 283–289.

Hiatt, J. B., R. P. Patwardhan, E. H. Turner, C. Lee, and J. Shendure, 2010 Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* 7: 119–122.

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

Karim, A. S., K. A. Curran, and H. S. Alper, 2013 Characterization of plasmid burden and copy number in *Saccharomyces cerevisiae* for optimization of metabolic engineering applications. *FEMS Yeast Res.* 13: 107–116.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, 2003 Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.

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

Kwasniewski, J. C., I. Mogno, C. A. Myers, J. C. Corbo, and B. A. Cohen, 2012 Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc. Natl. Acad. Sci. USA* 109: 19498–19503.

Lee, T. A., P. Jorgensen, A. L. Bognar, C. Peyraud, D. Thomas *et al.*, 2010 Dissection of combinatorial control by the Met4 transcriptional complex. *Mol. Biol. Cell* 21: 456–469.

Lengauer, C., K. W. Kinzler, and B. Vogelstein, 1998 Genetic instabilities in human cancers. *Nature* 396: 643–649.

Libkind, D., C. T. Hittinger, E. Valério, C. Gonçalves, J. Dover *et al.*, 2011 Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proc. Natl. Acad. Sci. USA* 108: 14539–14544.

Maclsaac, K. D., T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo *et al.*, 2006 An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7: 113.

Maerkl, S. J., and S. R. Quake, 2007 A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315: 233–237.

- Mai, B., and L. Breeden, 1997 Xbp1, a stress-induced transcriptional repressor of the *Saccharomyces cerevisiae* Swi4/Mbp1 family. *Mol. Cell. Biol.* 17: 6491–6501.
- McIsaac, R. S., A. A. Petti, H. J. Bussemaker, and D. Botstein, 2012 Perturbation-based analysis and modeling of combinatorial regulation in the yeast sulfur assimilation pathway. *Mol. Biol. Cell* 23: 2993–3007.
- Measday, V., K. Baetz, J. Guzzo, K. Yuen, T. Kwok *et al.*, 2005 Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc. Natl. Acad. Sci. USA* 102: 13956–13961.
- Melnikov, A., A. Murugan, X. Zhang, T. Tesileanu, L. Wang *et al.*, 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30: 271–277.
- Miller, A. W., C. Befort, E. O. Kerr, and M. J. Dunham, 2013 Design and use of multiplexed chemostat arrays. *J. Vis. Exp.* 72: 50262.
- Monteiro, P. T., N. D. Mendes, M. C. Teixeira, S. d'Orey, S. Tenreiro *et al.*, 2008 YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 36: D132–D136.
- Morris, D. R., and A. P. Geballe, 2000 Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* 20: 8635–8642.
- Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha *et al.*, 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344–1349.
- Patwardhan, R. P., J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith *et al.*, 2012 Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30: 265–270.
- Payen, C., S. C. Di Rienzi, G. T. Ong, J. L. Pogachar, J. C. Sanchez *et al.*, 2014 The dynamics of diverse segmental amplifications in populations of *Saccharomyces cerevisiae* adapting to strong selection. *G3 (Bethesda)* 4: 399–409.
- Payen C., A. B. Sunshine, G. T. Ong, J. L. Pogachar, W. Zhao *et al.*, 2015 Empirical determinants of adaptive mutations in yeast experimental evolution. *bioRxiv*: DOI: <http://dx.doi.org/10.1101/014068>.
- Petti, A. A., R. S. McIsaac, O. Ho-Shing, H. J. Bussemaker, and D. Botstein, 2012 Combinatorial control of diverse metabolic and physiological functions by transcriptional regulators of the yeast sulfur assimilation pathway. *Mol. Biol. Cell* 23: 3008–3024.
- Rhee, H. S., and B. F. Pugh, 2012 Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* 483: 295–301.
- Ronald, J., R. B. Brem, J. Whittle, and L. Kruglyak, 2005 Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* 1: e25.
- Schmieder, R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863–864.
- Shortle, D., P. Novick, and D. Botstein, 1984 Construction and genetic characterization of temperature-sensitive mutant alleles of the yeast actin gene. *Proc. Natl. Acad. Sci. USA* 81: 4889–4893.
- Siggers, T., M. H. Duyzend, J. Reddy, S. Khan, and M. L. Bulyk, 2011 Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* 7: 555.
- Skelly, D. A., J. Ronald, and J. M. Akey, 2009 Inherited variation in gene expression. *Annu. Rev. Genomics Hum. Genet.* 10: 313–332.
- Skelly, D. A., M. Johansson, J. Madeoy, J. Wakefield, and J. M. Akey, 2011 A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* 21: 1728–1737.
- Smith, F. W., M. J. Hawkesford, I. M. Prosser, and D. T. Clarkson, 1995 Isolation of a cDNA from *Saccharomyces cerevisiae* that encodes a high affinity sulphate transporter at the plasma membrane. *Mol. Gen. Genet.* 247: 709–715.
- Starita, L. M., J. N. Pruneda, R. S. Lo, D. M. Fowler, H. J. Kim *et al.*, 2013 Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. USA* 110: E1263–E1272.
- Stern, D. L., and V. Orgogozo, 2008 The loci of evolution: How predictable is genetic evolution? *Evolution* 62: 2155–2177.
- Struhl, K., 1989 Molecular mechanisms of transcriptional regulation in yeast. *Annu. Rev. Biochem.* 58: 1051–1077.
- Su, N.-Y., I. Ouni, C. V. Papagiannis, and P. Kaiser, 2008 A dominant suppressor mutation of the *met30* cell cycle defect suggests regulation of the *Saccharomyces cerevisiae* Met4-Cbf1 transcription complex by Met32. *J. Biol. Chem.* 283: 11615–11624.
- Teixeira, M. C., P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes *et al.*, 2006 The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 34: D446–D451.
- Teixeira, M. C., P. T. Monteiro, J. F. Guerreiro, J. P. Gonçalves, and N. P. Mira *et al.*, 2014 The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 42: D161–D166.
- Tirosh, I., S. Reikhav, A. A. Levy, and N. Barkai, 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324: 659–662.
- Titz, B., S. Thomas, S. V. Rajagopala, T. Chiba, T. Ito *et al.*, 2006 Transcriptional activators in yeast. *Nucleic Acids Res.* 34: 955–967.
- Wilkening, S., G. Lin, E. S. Fritsch, M. M. Tekkedil, S. Anders *et al.*, 2014 An evaluation of high-throughput approaches to QTL mapping in *Saccharomyces cerevisiae*. *Genetics* 196: 853–865.
- Wray, G. A., 2007 The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8: 206–216.
- Yona, A. H., Y. S. Manor, R. H. Herbst, G. H. Romano, A. Mitchell *et al.*, 2012 Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl. Acad. Sci. USA* 109: 21010–21015.
- Yun, Y., T. M. A. Adesanya, and R. D. Mitra, 2012 A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. *Genome Res.* 22: 1089–1097.
- Zhu, J., and M. Q. Zhang, 1999 SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15: 607–611.

Communicating editor: A. Gasch

GENETICS

Supporting Information

www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.188037/-/DC1

Comprehensive Analysis of the *SUL1* Promoter of *Saccharomyces cerevisiae*

Matthew S. Rich, Celia Payen, Alan F. Rubin, Giang T. Ong, Monica R. Sanchez,
Nozomu Yachie, Maitreya J. Dunham, and Stanley Fields

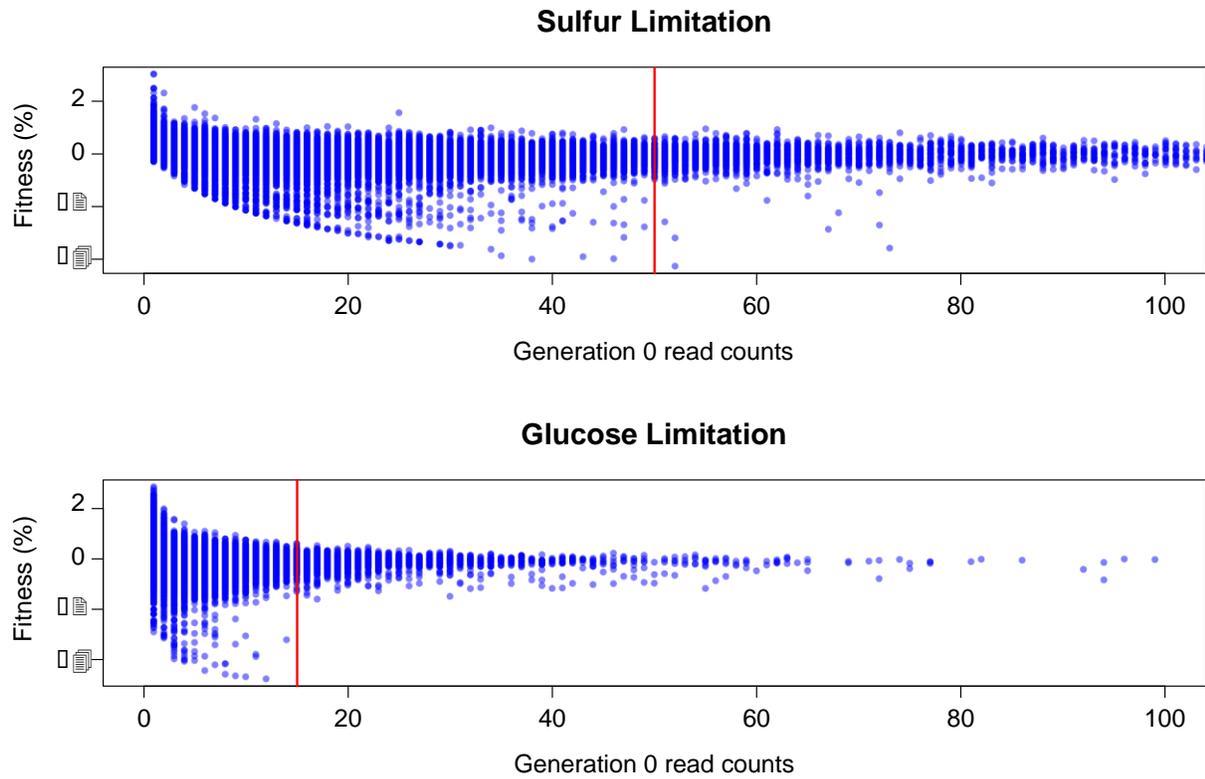


Figure S1 Determining minimum read count threshold for variant filtration. Barcodes linked to wildtype *SUL1* promoters were used to set a minimum generation 0 read count for variant filtration. The fitness of each wildtype barcode is plotted against that barcode's generation 0 read count. Filtration thresholds (red lines) were set at a point where the dependency between read count and barcode fitness was not apparent.

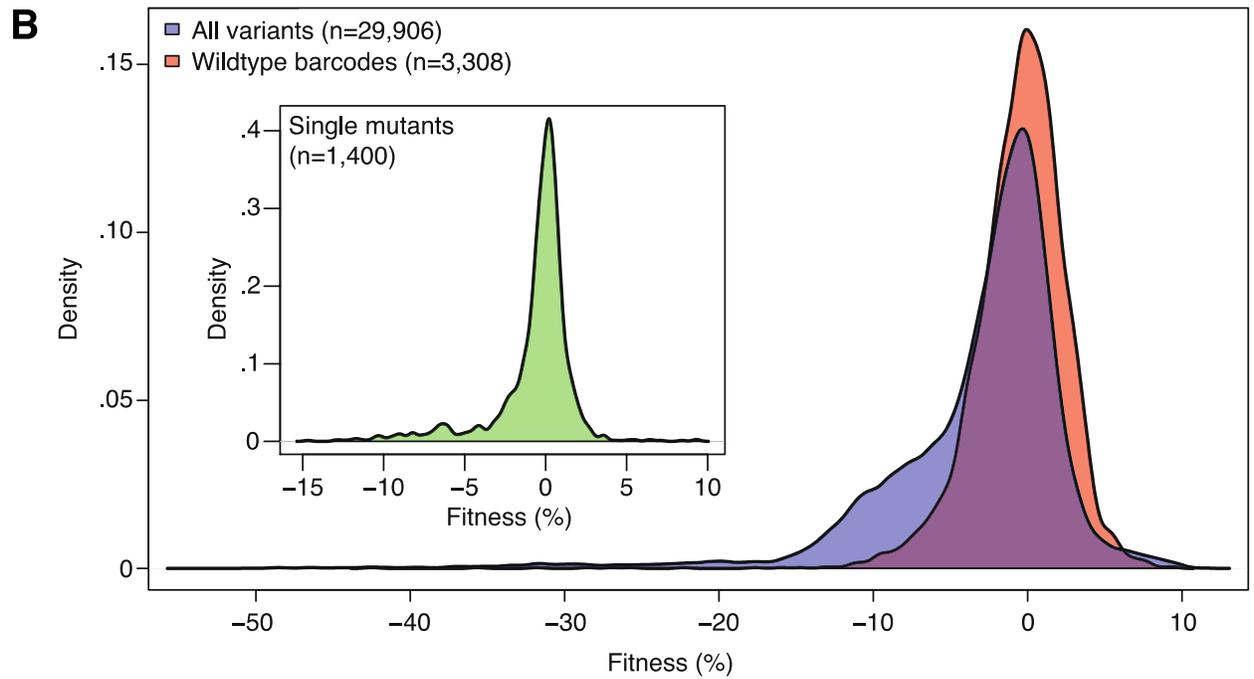
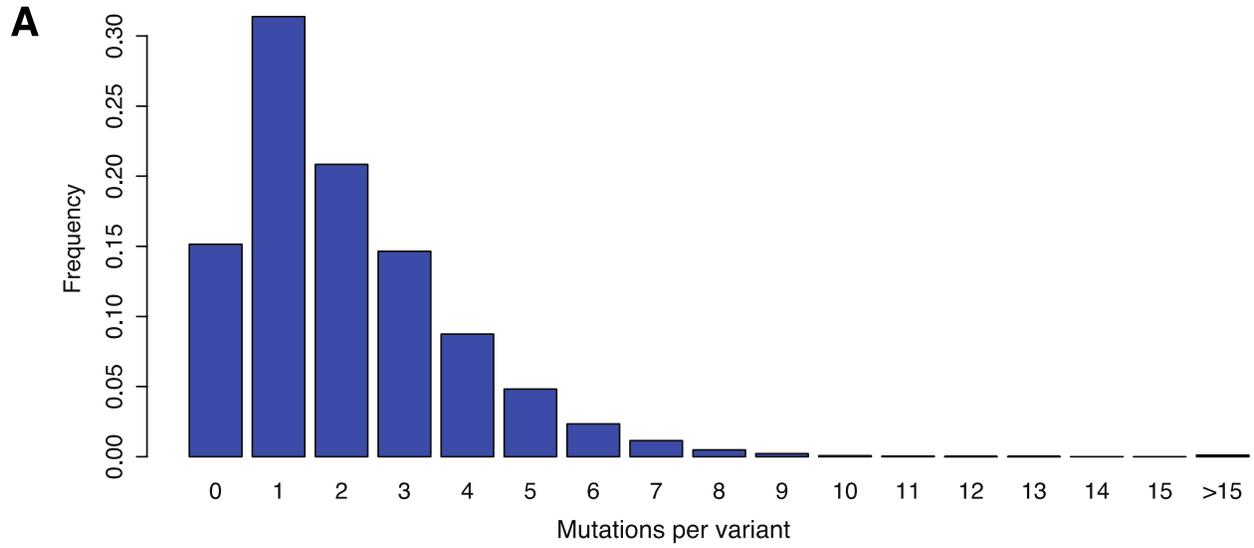


Figure S2 Variant distributions. (A) Histogram showing number of mutations per variant. (B) Density plot of variant fitness change distributions.

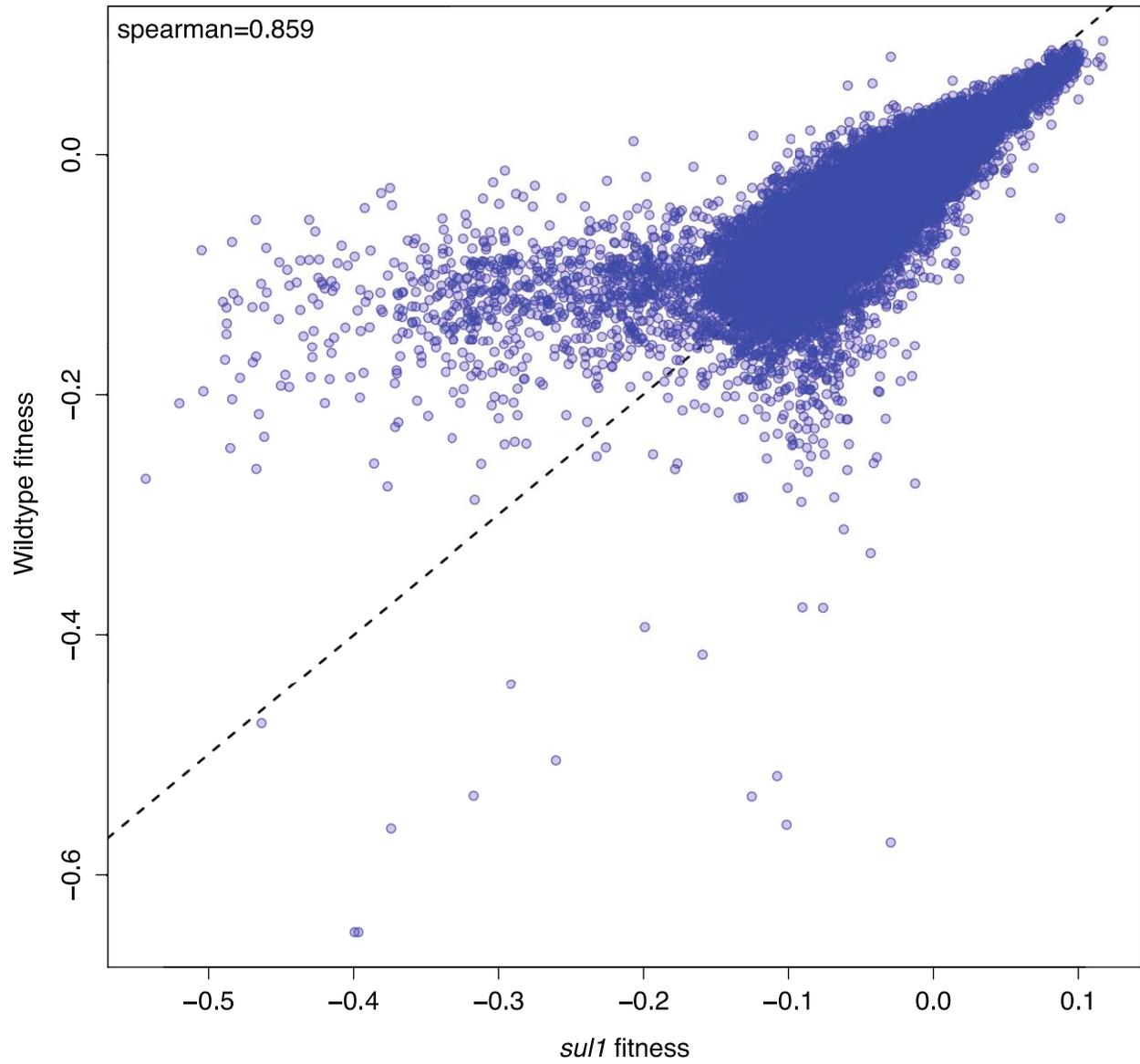


Figure S3 Correlation between fitness changes of variants in FY3 and FY3*sul1* Δ . Only variants passing filtration in both datasets (n=29,763) are shown. The dashed line shows $y=x$.

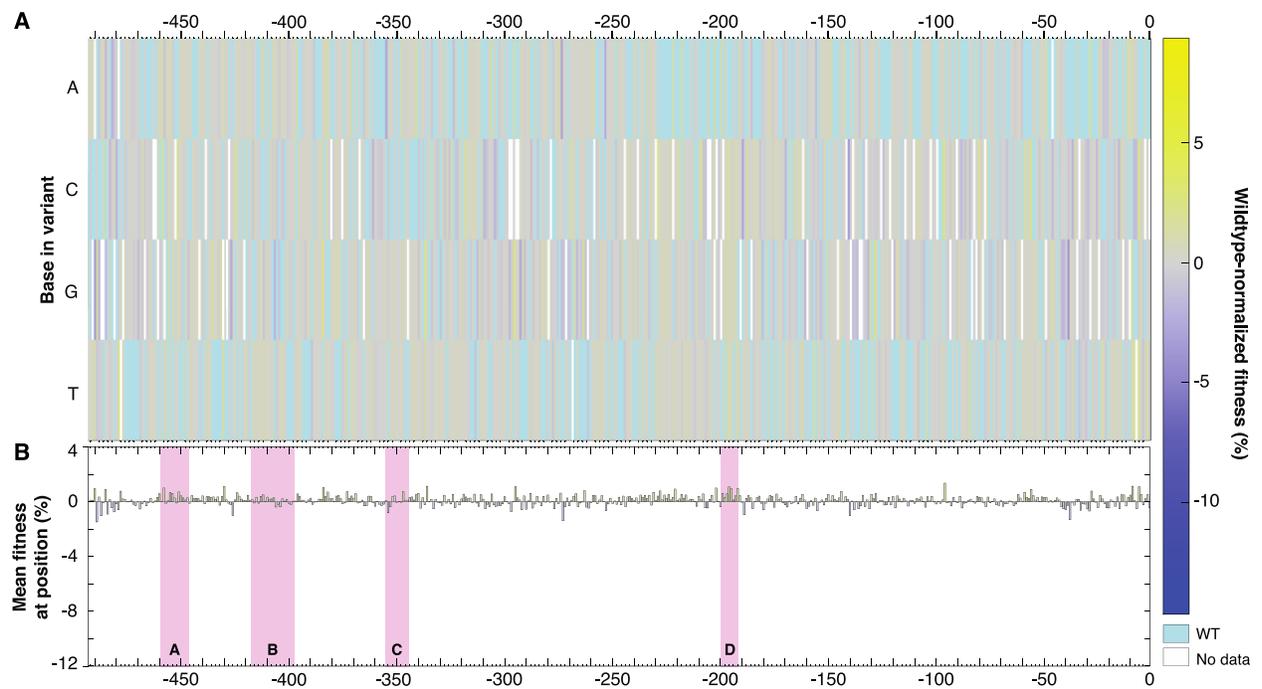


Figure S4 The effect of single mutations in *SUL1* promoter on glucose-limited fitness. (A) The fitness change of each single mutation in our dataset is plotted as a heatmap. At each position in the promoter along the x-axis, the base present in each single mutant is ordered on the y-axis, and every cell is shaded based on the fitness of that mutation. Missing data are white cells in the heatmap and the wildtype base at each position is teal. (B) Mean fitness change of all variants that are mutated at that position. Sites examined in Figure 3 are marked with pink backgrounds.

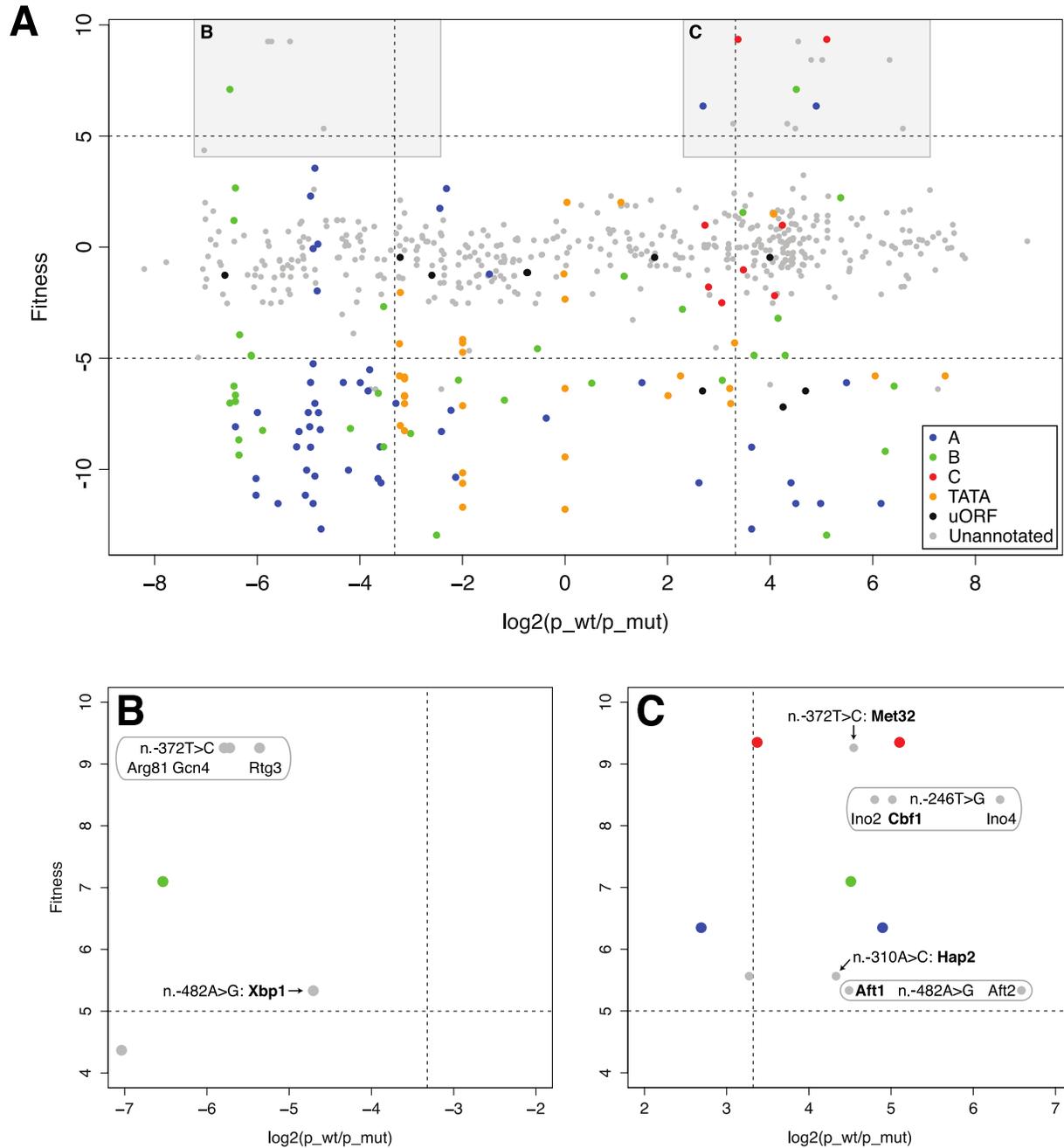


Figure S5 Identifying cryptic regulatory sites uncovered by mutations. (A) All motifs that strongly match either a wildtype or mutant 25-mer ($P < 0.001$). Motifs are plotted by the change to match significance of each mutation (x-axis) and the fitness of each single mutant (y-axis). Points are colored based on their annotated function. (B) Inset showing motif matches that are weakened by mutations that confer fitness increases. Unannotated mutations that match multiple motifs are grouped the transcription factor motif matched is noted. Transcription factors that we hypothesize to be functionally important at that site are in bold.

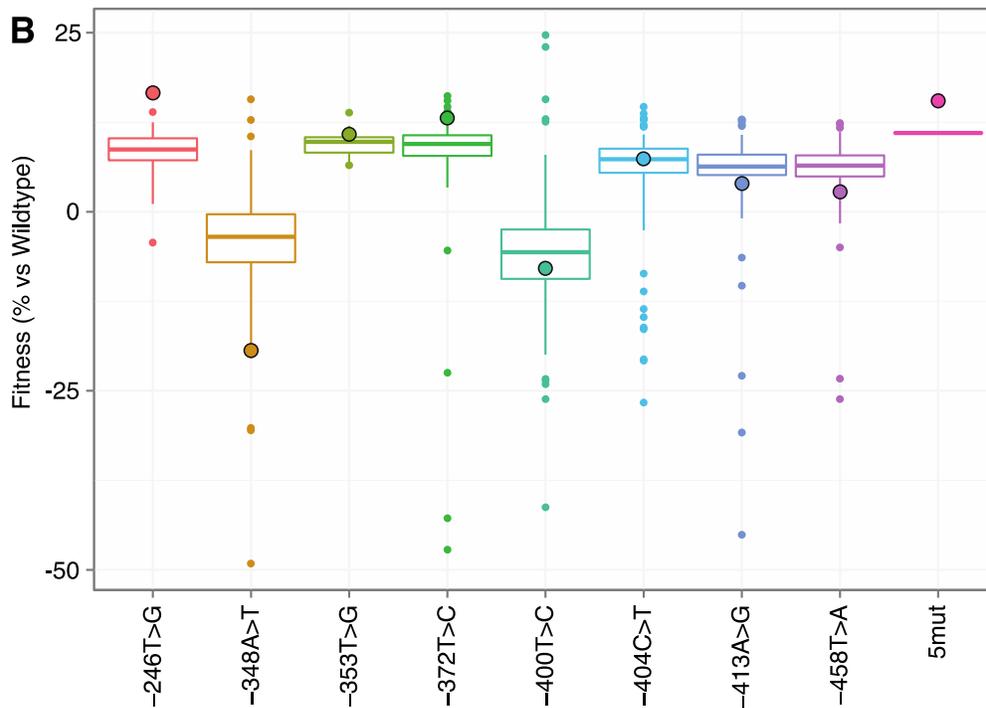
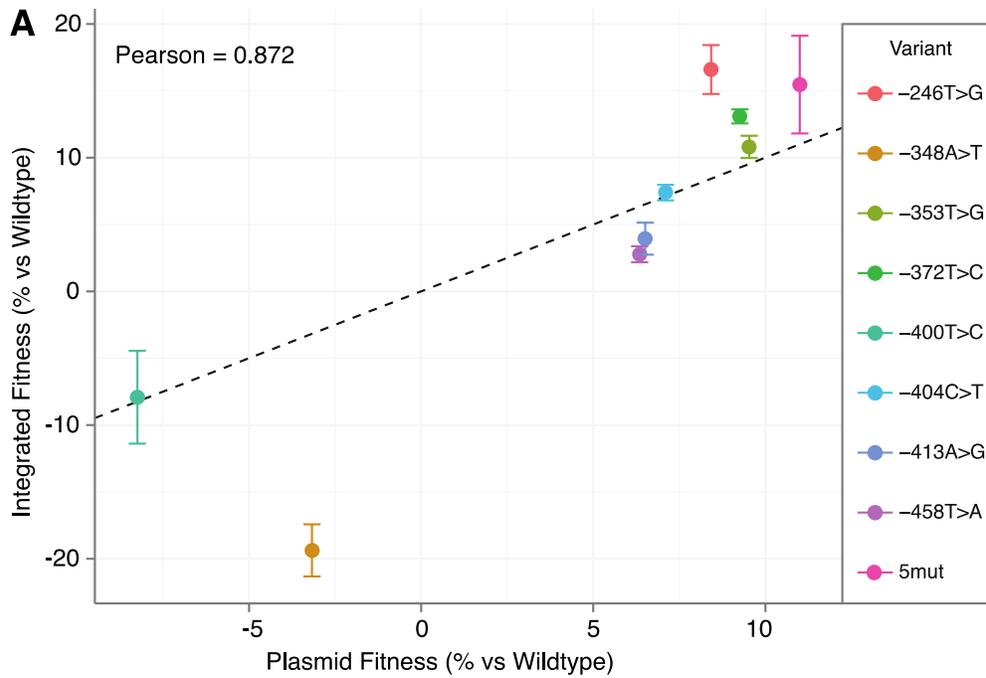


Figure S6 Correlation of fitness measurements between pooled plasmid assays and genomic Integrations. (A) Scatterplot showing fitness measurements for strains based either on a pooled assay in which promoter variants are found on centromeric plasmids (x-axis) or individual measurements in which promoter variants were integrated scarlessly at the *SUL1* promoter (y-axis). Error bars represent standard deviations ($n=2$). The dashed line shows $y=x$. (B) Boxplots showing the distribution of fitness values amongst redundantly barcoded variants tested in (A). Whiskers for boxplots span 95% of the total distribution. Larger outlined points mark the fitness change of strains where each variant has been integrated at the *SUL1* promoter. In both panels, fitness measurements are normalized to the wildtype fitness. The variant “5mut” comprises five mutations: -353T>G, -372T>C, -404C>T, -413A>G, and -458T>A.

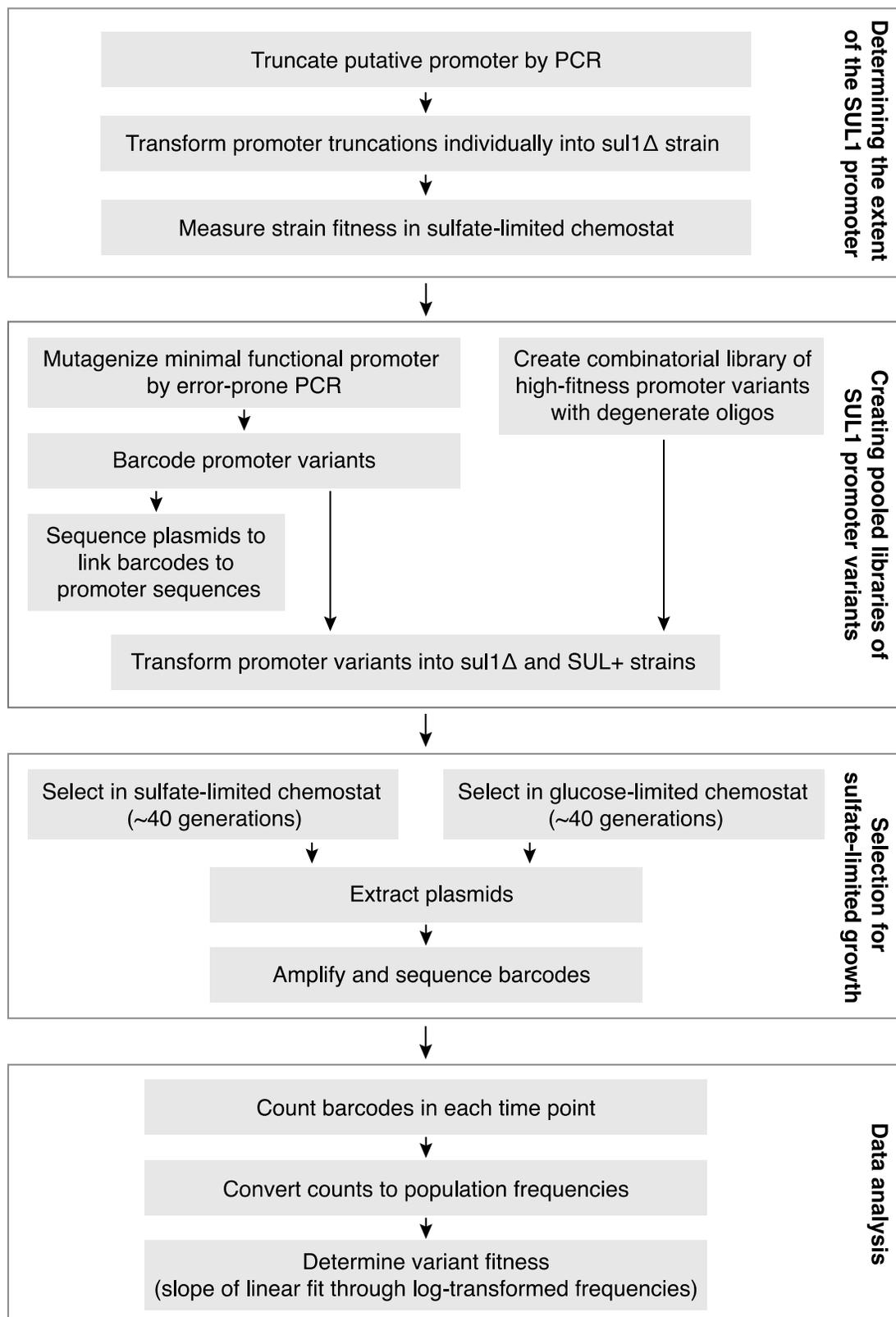


Figure S7 Flow-chart of experiments presented in this study

Table S1 Oligonucleotides used in this study.

Oligo#	Oligo Name	Oligo Sequence
266	pRS416-Sacl_chr2:792107_R	TCACTAAAGGGAACAAAAGCTGGATTCGGCTCAAGCTCAAGTTAACTTCTAGACC
283	pRS-Barcode@HindIII	GGCCCCCCTCGAGGTCGACGGTATCGATANNNNNNNNNNNAGCTTGATATCGGGATCC
294	pRS416-RI+HI_chr2:788548_F	ACGGTATCGATAAGCTTGATATCGGGATCCGACCGACAGTGCTGAACTGTGAG
295	pRS416-RI+HI_chr2:788742_F	ACGGTATCGATAAGCTTGATATCGGGATCCCGCCACCTCGAGTGCAC
296	fill-in_283_pMR002	CTCACAGTTCAGCACTGTCGGTCCGGATCCCGATATCAAGCT
297	SUL1p-mut_R	AGCATCCTCCTGATTATGCACATATTCAGTCGAGCTCTTACGTGACATATTTCTTCCGAG
314	pRS416-RI+HI_chr2:788643_F	GACGGTATCGATAAGCTTGATATCGGGATCCGTGAATAAGATGTGGCTGTGATAACC
315	pRS416-RI+HI_chr2:788845_F	GACGGTATCGATAAGCTTGATATCGGGATCCGCAAATCGGAATTTGAGTCACAGATC
316	pRS416-RI+HI_chr2:788943_F	GACGGTATCGATAAGCTTGATATCGGGATCCTTTGAACACTTCTACCTGTTTCATGTC
317	pRS416-RI+HI_chr2:789043_F	GACGGTATCGATAAGCTTGATATCGGGATCCGTGAAGTAAAATGTGTTGTAATGCACATGG
318	pRS416-RI+HI_chr2:789130_F	GACGGTATCGATAAGCTTGATATCGGGATCCGTGAAACCGATTATATAAAAGTATATTAGCTGAC
319	pRS416-RI+HI_chr2:789211_F	GACGGTATCGATAAGCTTGATATCGGGATCCCTGCAGAATACTCGGAAAGAAT
325	pRS416-HindIIIBC-seqF	GGGCCCCCCTCGAGGTCGACGGTATCGATA
326	pMR002-seqR	TTCCCTATATGTGCGGTATTCTGATTCAAATACTTCGATATCAGCATCCTCCTGATTATGCACATATTCAGTCGAGCTCTTACGTGACAT
327	P5-M13F	AATGATACGGCGACCACCGAGATCTACACTGTAAAACGACGGCCAGT
328	P7_SUL1-109_R	CAAGCAGAAGACGGCATAACGAGATCCTCATACCATTGTGAAGTCCGTCT
344	nexV2_SUL1-109_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCATACCATTGTGAAGTCCGTCT
380	SUL1p_6mut_F1	TCTCGCTGGARCCACAGTGYGGCTTTGCAATTTTCAAATCGGAATTTGAGYCACAGATCCCAGAAAAACKCCACACCTTCCCCACGCAG
381	SUL1p_6mut_R1	CACAGTGCATAATTTCAKATGACAGTGTTTCGAGAAAAGACATGAACAGG
382	SUL1p_6mut_F2	GTGCACTTTTTTAAATAAGATCWCGTGAATTGTCCAATCTGACTTTT
387	SUL1p_6mut_seqR	GGGTCTAATTTGACAACCTGTTTCGTTATCGCTTGCTGCGTGGGGAAGGTGTGG
395	P5_X_M13_F1	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNTGTAAAACGACGGCCAGT
442	SUL1p_6mut_seqF2	CGATAAGCTTGATATCGGGATCCCGCCACCTCGAGTGCACCTTTTTTAAATAAAGATC
443	SUL1p_6mut_seqI2	ATTAGACCCATAATAATTTTGAACACTTCTACCTGTTTCATGTCTTTTCTCGAACACT
517	n.-353T>G	AGATCCCAGAAAAACGCCACACCTTCCCCAC

518	n.-372T>C	TCGGAATTTGAGCCACAGATCCCAG
519	n.-246T>G	GTCTTTTCTCGAACACTGTCATGTGAAATTATGCA
520	n.-404C>T	GGAACCACAGTGTGGCTTTGCAATT
521	n.-413A>G	AGTCTCGCTGGAGCCACAGTGC GGC
522	n.-458T>A	GTGCACTTTTTTAATAAAGATCACGTGTAATTGTC
603	-33A>G	ggtcaatagctttaaaataGaaataaatccctgcag
604	-400T>C	cagtgcggcCttgcaattttgc
605	-348A>T	gaaaaactccacTccttccccac
678	pSUL1::URA3_F	TCCCCCACTTTTTGTGTTACCGCCACCTCGAGTGACGgattcgtaatctccgaacag
679	pSUL1::URA3_R	ACATATTCAGTCGAGCTCTTACGTGACATgggtaataactgatataattaattgaagct
680	pSUL1::URA3_F2	CACAAACCAATGAAAAAAGGCTTCAGAAAACTTGCTGTAAAGTCCCCCACTTTTTTG
681	pSUL1::URA3_R2	attcaataactcgcgatcagcatcctcctgattatgcacatattcagtcgagctcttac
685	pSUL1_687bp_F	gaccgacagtgtgaactgtgag
PGO40	SUL1::NatMX4 Forward primer	TTGAAACTAGATATTCTACTTGACACTAACTTTTTTTGTATTTCTGGACAGATTGTACTGAGAGTGCACCATA
PGO41	SUL1::NatMX4 reverse primer	TGAAAACGGCCCTTTTCGTGTACTAATGCTTTCAGGAAATTATAAAATTCCTTACGCATCTGTGCGGTATTTCA
NexV2ad2_N ²	NexV2ad2_N ²	CAAGCAGAAGACGGCATAACGAGATNNNNNNNNNGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
M13F	M13F	TGAAAACGACGGCCAGT
P5	P5	AATGATACGGCGACCACCGAGATCTACAC

1 – P5_N_M13_F is a set of indexed oligonucleotides used to demultiplex Illumina sequencing datasets (by the second index read.) Each oligonucleotide has a unique sequence in the region marked by Ns.

2 - NexV2ad2_N is a set of indexed oligonucleotides used to demultiplex Illumina sequencing datasets (by the first index read). Each oligonucleotide has a unique sequence in the region marked by Ns.

Table S2 Yeast strains used in this study.

Name	Mating type	Genotype	Reference
FY2	alpha	ura3-52	(Winston <i>et al.</i> 1995)
FY3	a	ura3-52	(Winston <i>et al.</i> 1995)
FY4	a	prototroph	(Winston <i>et al.</i> 1995)
YMD3017	a	SUL1::NatMX4, ura3-52 Gal+	This study
YMR002	a	FY3 P _{SUL1} ::URA3 ura3-52	This study
YMD3018	alpha	sul1::NatMX	This study
YMR008	a	P _{SUL1} -246T>G ura3-52	This study
YMR009	a	P _{SUL1} -404C>T ura3-52	This study
YMR010	a	P _{SUL1} -458T>A ura3-52	This study
YMR011	a	P _{SUL1} -400T>C ura3-52	This study
YMR012	a	P _{SUL1} -348A>T ura3-52	This study
YMR017	a	P _{SUL1} -353T>G ura3-52	This study
YMR018	a	P _{SUL1} -372T>C ura3-52	This study
YMR019	a	P _{SUL1} -413A>G ura3-52	This study
YMR020	a	P _{SUL1} -5mut ura3-52	This study
YMR022	a	P _{SUL1} -246T>G	This study
YMR024	a	P _{SUL1} -458T>A	This study
YMR025	a	P _{SUL1} -400T>C	This study
YMR026	a	P _{SUL1} -348A>T	This study
YMR027	a	P _{SUL1} -353T>G	This study
YMR028	a	P _{SUL1} -372T>C	This study
YMR029	a	P _{SUL1} -413A>G	This study
YMR030	a	P _{SUL1} -5mut	This study
YMD1214	a	ho::kanMX-eGFP	(Payen <i>et al.</i> 2014)
YMD1139	a	pFA6a-TEF2Pr-eGFP-ADH1-Primer-NATMX4	D. Breslow

File S1: Software used for data analysis. (.zip, 3 MB)

Available for download as a .zip file at:

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.188037/-/DC1/FileS1.zip>

File S2: Raw data files. (.zip, 44 MB)

Available for download as a .zip file at:

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.188037/-/DC1/FileS2.zip>