Online methods

Library design. We designed a total of 6500 promoters, divided into several sets of promoters, each aimed at examining a specific aspect of transcriptional regulation. We designed the library using a "lego"-like methodology: each promoter is composed from a "context" DNA sequence and several regulatory elements, such as TF binding sites and nucleosome disfavoring sequences. With a few exceptions (described below), the elements are integrated into the promoter by replacing the background DNA at the relevant position. Since we were interested in the position of the regulatory elements relative to the YFP gene, our coordinates are specified from the 3' end of the oligo to the 5' end and are 1-based. Correspondingly, the start of each element is its 3' end position. In several cases, regulatory elements were integrated by insertion. In these cases, the insertion position is described in the library design file by a non-integer coordinate. The integer portion of the number specifies the number of background nucleotides downstream to the insertion while the fraction indicates the order of the inserted element from 3' to 5' (in case several elements were inserted into a similar position). When regulatory elements were integrated into the background sequence by insertion rather than by replacement, the sequence is truncated in its 5' to maintain uniform oligo lengths. The library description file contains a unique identifier for every promoter, a description text that specifies the background DNA (referred to as "Context" DNA), regulatory elements, barcode and PCR primers sequence and position (for technical reasons, the "-1" coordinate indicates that the element does not exist in the promoter), and the library oligo sequence.

Background sequences. The library is constructed with two primary background DNA sequences, termed mGal1-10 (in the description files: "GAL1-10_NULL") and mHis3 (in the description files: "HIS3_NULL"). The mGal1-10 background is a sequence from the native yeast Gal1-10 promoter in which the known regulatory elements were mutated. The mHis3 is taken from the yeast His3 promoter with known regulatory elements deleted. Coding ("CDS"), intergenic and promoter backgrounds were selected randomly from the corresponding region in yeast genome as annotated in the UCSC genome browser[36]. The "GC40Random*" backgrounds were generated randomly with mean 40% dG/dC. "<gene name>_NATIVE" are unmodified DNA sequences from the specified gene promoter[36]. The "GAL1_10_GINIGER*" sequences are variations of the yeast Gal1-10 promoter.

Regulatory elements. For each TF, we selected one or more binding sites based on the literature cited in the regulatory elements description file. Our regulatory elements

also contain nucleosome disfavoring sequences (indicated by type: “Boundary”). The main nucleosome disfavoring sequence that was used in the library is a stretch of poly(dT) of length 15bp.

Design of promoter barcodes. The 5’ edge of each designed promoter contains a unique 10bp sequence, which we use as a barcode for the promoter that allows us to uniquely identify it using a short sequencing read. We designed the barcode sequences such that every pair of barcodes differs from each other in at least 3bps, allowing us to correctly identify every barcode even if it contains a single basepair mutation. We excluded low complexity sequences from being barcodes as these may affect the PCR reactions. To minimize the effect of the barcode on the promoter expression, we also excluded sequences that resemble any known yeast transcription factor binding site sequence, by ensuring that they do not match known published consensus sequences[20,37].

Design of expression bin barcodes. To identify the expression bin of each promoter from the multiplexed sequencing land, we added a unique 5bp tail to the 5’ end of the primer that is used to amplify the synthetic promoter region of every promoter expression bin. We chose the bin barcodes such that every pair of bin barcodes differ from each other in the last two basepairs, such that a single basepair mutation cannot cause an error in the bin mapping. We excluded low complexity sequences from being barcodes as these may affect the PCR reactions.

Synthetic promoter library generation. To generate a large-scale library of 6500 different types of promoters each fused to a fluorescent reporter, with each cell containing one of the 6500 synthetic promoters, we first used Agilent Oligo library synthesis (OLS) technology to produce a pool of 6500 different single-stranded 150-mers. The library is synthesized on programmable microarrays by Agilent[14,38] and then provided to us as an oligo pool in a single tube (10pmol). Each oligo contains common priming sites and restriction sites at both ends, as well as a 10bp unique barcode, leaving 103bp for the variable promoter region in which we performed the various library manipulations. The pool was dissolved in 200µl TE. We divided 12ng (0.026%) of the library into 96 wells and amplified each well using PCR, where the volume of each PCR reaction was 50µl. Each reaction contained 24µl of water containing 0.125ng DNA, 10µl of 5X HerculaseII reaction buffer, 10µl 2.5mM dNTPs mix, 2.5µl 20uM 5’ primer, 2.5µl 20uM 3’ primer, 1µl Herculase II Fusion DNA Polymerase (Agilent[2]). The parameters for PCR were 95°C for 1 min, 12 cycles of 95°C for 20s and 68°C for 1 min, each, and finally one cycle of 68°C for 4 min. The primers used to amplify the library were GGGGACCAGGTGCCGTAA (forward

primer) and TGATCGCCCTAGGATCGC (reverse primer). The PCR products from all 96 wells were joined and concentrated using Amicon Ultra, 0.5ml 30K centrifugal filters for DNA Purification and Concentration. The concentrated DNA was then purified using a PCR minielute purification kit (Qiagen) according to the manufacturer's instructions.

Construction of the library plasmid. Plasmid pKT103[39], which contains yEVenuse, was used as a vector backbone to create the recipient plasmid (pPAL1_His3core) for the library. The KAN cassette was replaced with a NAT cassette. We amplified TEF2promoter-mCherry from pAG60-TEF2-Cherry plasmid[40] and inserted as EcoO109I/AatII fragment into pKT103. We amplified URA3 from pRS316 and inserted it to the recipient plasmid. In order to allow for chromosomal segregation of the library plasmids we introduced an ARS-CEN sequence. The ARS-CEN sequence was amplified from pRS316 plasmid and inserted as EcoRI/EcoRV fragment. We amplified the first 100bp upstream to genomic HIS3 ATG that served as a core promoter for the library, and inserted it upstream of the yEVenus ATG. The HIS3 core promoter was flanked by a sequence with two restriction sites that were used for the ligation of the 150bs pooled library.

Ligation and transformation. Purified library DNA (150ng) was cut with the restriction enzymes AvrII and SexAI (Fermentas) for 2 hours at 37°C in a reaction mixture containing 3µl FD buffer, 0.8µl of each enzyme, and 25.4µl DNA. Digested DNA was separated from smaller fragments by electrophoresis on a 2.5% agarose gel stained with Ethidium Bromide. Only right size fragments (128bp + 9bp overhang) were cut from the gel and eluted using electoelution Midi GeBAflex tubes. Eluted DNA was precipitated using standard NaAcetate\Isoprpoanol protocol. To prepare the vector for cloning, the plasmid was cut with the restriction enzymes AvrII and SexAI for 2 hours at 37°C in a reaction mixture containing 6µl FD buffer, 3µl of each enzyme, 3µl Alkaline Phosphatase (Fermentas), and 3ug of the plasmid in a total volume of 60µl. Digested DNA was purified using a PCR purification kit (QIAGEN). The digested plasmid DNA library were ligated for 1/2 hr at room temperature in a 10µl reaction that contained 150ng plasmid, 2.8ng library (molar ratio of 1:1), 1µl CloneDirect 10X Ligation Buffer, and 1µl CloneSmart DNA Ligase (Lucigen Corporation). The ligation in the plasmid was upstream of a YFP reporter gene and a short core promoter (100bp TATA containing core promoter taken from the native His3 yeast promoter). Ligated DNA was transformed into 7 tubes (containing 25µl, each) of E. cloni 10G electrocompetent cells (Lucigen) which were then plated on 28 LB/amp 15cm plates. Sixteen hours after transformation, the plates containing

50,000 colonies each were scraped into LB medium and the plasmid was purified using a plasmid maxi kit (QIAGEN). To minimize the number of plasmids with multiple inserts, we cut the plasmid only with the SexAI restriction enzyme, ran it on a gel stained with crystal violet, purified the digested plasmid from gel with QIAquick gel purification kit (QIAGEN), and then religated the plasmid using CloneSmart DNA Ligase. Transformation was performed as described above. Cells from 6 transformations were plated on 30 15cm plates. Colonies were scraped (84,000 per plate) and plasmids were purified with a plasmid maxi kit (QIAGEN).

Yeast transformation. To achieve a high efficiency of transformation, we used the electroporation protocol[3] to transform the library plasmid into Y8205 strain (kind gift from C. Boon). For a single transformation we grew the cells culture to OD^{600} 1.4, 10^8 cells were washed twice with cold DDW and once with cold 1M sorbitol. Cells were suspended in 40µl cold 1M sorbitol. 0.5ug of library plasmid were mixed with the cells. The mixture was kept on ice for 5 min and transferred to an electroporation cuvette (0.2mm) and a 1.5 kV, 25 µF, 200Ω pulse was applied (Gene Pulser, Bio-Rad, Richmond, California, USA). The cells were diluted immediately with 0.95 ml of SCD-URA medium and later on 1:20, to OD_{600} 1, and transformed cells (120,000 transformants) were grown at 30°C for 72 hours until the culture reached stationary phase. To preserve the diversity of the library, we performed this transformation 23 times.

Sorting by Flow Cytometry. To adapt cells to the medium of the sorting, stationary yeast library cells were diluted in SC-Gal medium without amino acids except for histidine and leucin and were grown to stationary phase. Next, this culture was diluted again in similar medium (to OD_{600} 0.03-0.05), and grown to mid-exponential phase (OD_{600} 0.5-1.5) for sorting. Sorting was performed with the FACSAria cell sorter (Beckton-Dickinson) at the low sample flow rate and a sorting speed of no more than 20,000 cells/sec. In order to reduce the effect of extrinsic noise on promoter activity, we sorted only cells that were gated to have relatively homogeneous size and mCherry fluorescence (corresponding to ~1-2 plasmid copies). The library was sorted using two strategies. In one strategy (replicate 1), the cells were sorted three times recursively into four bins, producing a total of 64 bins (i.e., we first sorted all cells into four bins, applied the same sorting procedure to the cells of each bin, and again to the cells of each bin sorted at this second level). In another strategy (replicate 2), we sorted the cells directly into 16 bins. We chose these numbers of bins as a compromise between sorting time and having enough bins to resolve the expression level to a good resolution (demonstrated by the excellent agreement of R^2=0.99 that we obtained between our expression level

measurements and those of isolated strains). In both strategies, cells were sorted according to the ratio of YFP and mCherry (located in the plasmid with a promoter that is constant across the library, the yeast TEF2 promoter) , thereby normalizing for extrinsic noise effects. Since the mCherry reporter is driven by the same promoter across the entire library, this sorting by the ratio corrects for extrinsic noise across the cell population. The distribution of mCherry intensities peaked at several discrete levels, corresponding to the different number of plasmids integrated in each cell. The first peak was significantly higher than all others, suggesting that the vast majority of cells had a single plasmid (likely due to our use of low copy number plasmids and low concentration of plasmid DNA during the transformation process) and we thus gated the cells by the mCherry levels of the first peak to enrich for this population. In each sorting strategy, the expression bins contained equal fractions of the library cells, and we collected a total of 8,000,000 cells. We note that since both YFP and mCherry are stable and long-lived[19], their levels integrate the expression over several cell cycles, and thus, promoters that are expressed only in a specific cell cycle phase will have lower expression than those that are constitutively expressed throughout the cell cycle. For these reasons, our current paper made minimal use of cell-cycle dependent regulators.

Isolating control strains. To obtain isolated strains from the library, a small aliquot of cells from each bin was plated on SCD-URA plates. Six colonies from each bin were picked and further grown in a selective medium, and used for both sequencing and measurements. The PCR product containing the promoter from 96 individual colonies (3 from each bin) was sent to Sanger sequencing. Promoter activities from all sequenced stains were measured using both Flow Cytometry (with an "LSRII" cell analyzer and a similar protocol to the library measurement) and a plate reader (Tecan Infinite F500) as previously described[34].

Preparing sample for sequencing. Sorted cells were grown in 5ml SCD-Ura medium to stationary phase. One million cells from each bin were taken for colony PCR using specific primers corresponding to the promoter region of the plasmid. The 3' primer was common to all bins (5'-NNNNNTTATGTGATAATGCCTAGGATCGC-3', where "N"s represent random nucleotides). The 5' primer had a common sequence and a unique upstream 5bp barcode sequence (underlined) specific to each bin (5'-XXXXXGGGGACCAGGTGCCGTAAG-3', the "X"s represent the bin unique sequence). In replicate 2, DNA from each bin was subjected to PCR with two different 5' primers. PCR products were purified using ZR-96 DNA clean and Concentrator-5 (Zymo Research). Equal amount of DNA from each PCR product (2.35ng) were joined to one tube, concentrated using Amicon Ultra 0.5ml 30K

centrifugal filters and the sample (100ng) was then sent to sequencing by the SOLid system.

Sequencing. We sequenced the sample using Applied Biosystem's SOLiD high throughput sequencing platform (AB SOLiD) and mapped it to a reference sequence set which contained all combination of strain and expression bin barcode sequences using SHRiMP[3]. We obtained ~13M (replicate 1) and ~14M (replicate 2) uniquely mapped reads that each contain a pair of strain and expression bin barcode.

Deriving mean promoter expression levels. We first obtained the mean and standard deviation of the expression of each sorted flow cytometer bin, which we estimated from the expression levels of the cells sorted into each bin. Next, for each promoter, we estimated the fraction of its cells in each given bin based on the number of sequence reads that mapped for it in that bin (the reads of each bin were first normalized to match the fraction of the bin in the entire population). Finally, we defined the mean expression of each promoter as a weighted average of the mean expression of all bins, where the weight of each bin is the fraction of the promoter in that bin. Similarly, we computed the standard deviation of each promoter using the standard deviation of each bin and the distribution of the promoter across the bins. When our experiment consists of more than one round of sorting, we used the results of the last round as describe above.

Deriving expression predictions from in-vitro binding affinities. To compare our expression measurements of the effects of mutations to Gcn4 binding sites (**Fig. 4B**) to in vitro binding affinities of Gcn4 sites, we used a simple model to derive expression predictions from the in vitro affinities. To this end, we assumed that expression is linearly related to the promoter occupancy of Gcn4 occupancy, and computed Gcn4 occupancy as a function of the binding site dissociation constant (K_d) using a simple Michaelis-Menten based model, while adding a term that accounts for the extra observed activity of Gcn4 sites in reverse orientation. Thus, given the measured dissociation constant of a Gcn4 site, the predicted expression is given by:

$$expression = a * (1 + b * \delta) * \frac{[Gcn4]}{[Gcn4] + K_d}$$

where a, b, and [Gcn4] are the free parameters of the model (scaling factor, orientation factor, and Gcn4 concentration, respectively) and δ is an indicator function for whether the Gcn4 site is in reverse orientation relative to His3 native Gcn4 site, and we fit the free parameters so as to best fit our measured expression levels.

Visualizing comparisons of expression across promoter sets. Rather than using

traditional box-plots, we used the Matlab function “notBoxPlot”, written by Ron Campbell, to visualize various comparisons of expression values across different promoter sets.

Statistical analysis. To ensure that our various results are robust to outliers, we adapted robust statistical practices and used median instead of the mean as an estimator of the central tendency, and median absolute deviation (MAD) scaled to the normal distribution dispersion as an estimator of the standard deviation[41]. To assess the effect of various nucleosome disfavoring elements on expression, we used a two sided two-sample t-test. To show that the effect of surrounding sequence on the activity of regulatory elements is significantly larger than the barcode effect on strains with identical promoter, we used a two-sample F-test for equal variance between every two promoter groups (using the Matlab function “vartest2”).

Figure legends

Figure 1. Obtaining accurate expression measurements for thousands of designed promoter sequences. **(A)** Illustration of our experimental method. **(B)** Our method obtains highly reproducible expression measurements. Shown is a comparison of expression measurements (log-scale) obtained for two independent replicates done using two different cell sorting strategies (y-axis, replicate 1 sorted into 64 bins; x-axis, replicate 2 sorted into 16 bins, see Methods), along with lines (green) that correspond to a difference of 30% from the mean of the two replicates. 114 (1.75%) of the 6500 promoters that we designed fell outside the green lines and were filtered out from our analyses. **(C)** Barcodes have little effect on our expression measurements. Shown is the distribution of sequencing reads across the expression bins that we obtained for four pairs of promoters that differ only in their barcode sequence. See **Fig. S2** for 14 additional such promoter pairs. **(D)** Similar to (C), but for four sets of promoters where each set contains 10 (columns 3-4) or 20 (columns 1-2) promoters that differ only in their barcode sequence. For each set, shown are the individual expression measurements (gray dots), and their median (red line), standard error (orange bar), standard deviation (blue bar), and coefficient of variation (CV, standard deviation divided by the mean). **(E)** Our method obtains highly accurate expression measurements. We isolated 92 individual strains from our pool of transformed yeast cells and sequenced each of them to reveal their identity. Shown is a comparison of the expression for these strains when each strain was measured in isolation using a flow cytometer (x-axis) or within a single experiment using our method (y-axis).

Figure 2. Profiling the activity of most yeast transcription factors. **(A)** Consensus binding sites for 75 yeast transcription factors were separately inserted in their two possible orientations at the same position within a fixed promoter context (bottom illustration). Shown is a ranking of the resulting expression levels for each promoter, with the two site orientations of each TF colored red and green depending on whether they correspond to the orientation with higher or lower expression, respectively. For brevity, individual measurements for promoters with intermediate expression levels are not given (TF sites and their internal ranking are indicated in the box). Cyan and purple asterisks correspond to TFs with literature-reported activating or repressive roles, respectively. A horizontal black line marks the expression of the same fixed promoter above but without any known TF binding site,