Article

The evolution, evolvability and engineering of gene regulatory DNA

https://doi.org/10.1038/s41586-022-04506-6

Received: 8 February 2021

Accepted: 2 February 2022

Published online: 09 March 2022

Check for updates

Eeshit Dhaval Vaishnav[1,2,12]✉, Carl G. de Boer[3,4,12]✉, Jennifer Molinet[5,6], Moran Yassour[4,7,8], Lin Fan[2], Xian Adiconis[4,9], Dawn A. Thompson[2], Joshua Z. Levin[4,9], Francisco A. Cubillos[5,6] & Aviv Regev[4,10,11]✉

Mutations in non-coding regulatory DNA sequences can alter gene expression, organismal phenotype and fitness[1–3]. Constructing complete fitness landscapes, in which DNA sequences are mapped to fitness, is a long-standing goal in biology, but has remained elusive because it is challenging to generalize reliably to vast sequence spaces[4–6]. Here we build sequence-to-expression models that capture fitness landscapes and use them to decipher principles of regulatory evolution. Using millions of randomly sampled promoter DNA sequences and their measured expression levels in the yeast *Saccharomyces cerevisiae*, we learn deep neural network models that generalize with excellent prediction performance, and enable sequence design for expression engineering. Using our models, we study expression divergence under genetic drift and strong-selection weak-mutation regimes to find that regulatory evolution is rapid and subject to diminishing returns epistasis; that conflicting expression objectives in different environments constrain expression adaptation; and that stabilizing selection on gene expression leads to the moderation of regulatory complexity. We present an approach for using such models to detect signatures of selection on expression from natural variation in regulatory sequences and use it to discover an instance of convergent regulatory evolution. We assess mutational robustness, finding that regulatory mutation effect sizes follow a power law, characterize regulatory evolvability, visualize promoter fitness landscapes, discover evolvability archetypes and illustrate the mutational robustness of natural regulatory sequence populations. Our work provides a general framework for designing regulatory sequences and addressing fundamental questions in regulatory evolution.

Changes in *cis*-regulatory elements (CREs) have a major role in the evolution of gene expression[1]. Mutations in CREs can affect their interactions with transcription factors, change the timing, location and level of gene expression and affect organismal phenotype and fitness[2,3]. Although transcription factors evolve slowly because they each regulate many target genes, CREs evolve much faster and are thought to drive substantial phenotypic variation[7]. Thus, understanding how variation in *cis*-regulatory sequences affects gene expression, phenotype and organismal fitness is fundamental to our understanding of regulatory evolution[2].

A fitness function maps genotypes (which vary through mutations) to their corresponding organismal fitness values (where selection operates)[8]. A complete fitness landscape[9] is defined by a fitness function that maps each sequence in a sequence space to its associated fitness, coupled with an approach for visualizing the sequence space. Partial fitness landscapes have been characterized empirically[4,5,10], often defining fitness as the maximum growth rate of single-cell organisms[4,11]. Many empirical fitness landscape studies of proteins[12], adeno-associated viruses[13], catalytic RNAs[14], promoters[15] and transcription factor binding sites[16] have favoured molecular activities as fitness proxies because they are less susceptible to experimental biases and measurement noise[17]. In particular, the molecular activity of a promoter sequence as reflected in the expression of the regulated gene has been used to build a 'promoter fitness landscape'[18]. However, despite advances in high-throughput measurements, empirical fitness landscape studies often sample sequences in the local neighbourhood of natural ones and thus remain limited to a tiny subset of the complete sequence space, the size of which grows exponentially with sequence length (4^L for DNA or RNA, where L is the length of sequence)[4–6].

Understanding the relationship between promoter sequence, expression phenotype and fitness would allow us to answer fundamental

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]School of Biomedical Engineering, University of British Columbia, Vancouver, British Columbia, Canada. [4]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Departamento de Biología, Facultad de Química y Biología, Universidad de Santiago de Chile, Santiago, Chile. [6]ANID—Millennium Science Initiative Program, Millennium Institute for Integrative Biology (iBio), Santiago, Chile. [7]Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel. [8]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. [9]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [11]Present address: Genentech, South San Francisco, CA, USA. [12]These authors contributed equally: Eeshit Dhaval Vaishnav, Carl G. de Boer. ✉e-mail: edv@mit.edu; carl.deboer@ubc.ca; aviv.regev.sc@gmail.com

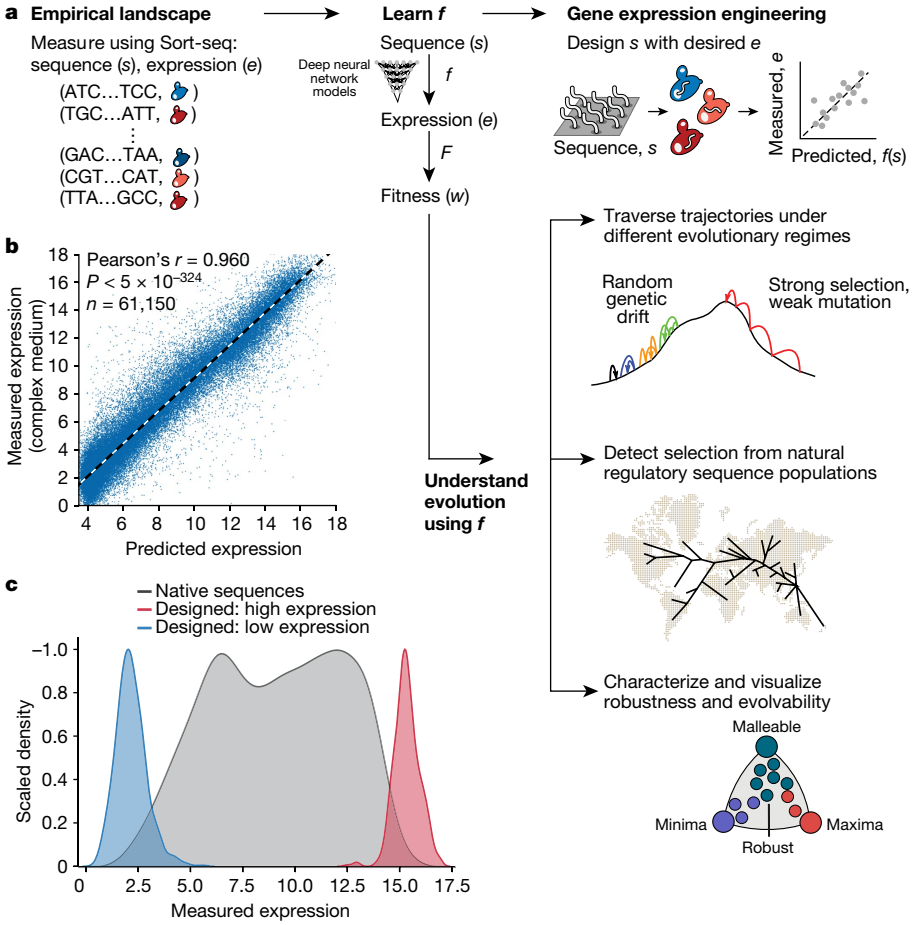


Fig. 1 | The evolution, evolvability and engineering of gene regulatory DNA. **a**, Project overview. **b**, Prediction of expression from sequence using the model. Predicted (*x* axis) and experimentally measured (*y* axis) expression in complex medium (YPD) for native yeast promoter sequences. Pearson's *r* and associated two-tailed *P* values are shown; dashed line: line of best fit. **c**, Engineering extreme expression values beyond the range of native sequences using a genetic algorithm and the sequence-to-expression model. Normalized kernel density estimates of the distributions of measured expression levels for native yeast promoter sequences (grey), and sequences designed (by the genetic algorithm) to have high (red) or low (blue) expression.

questions[6] in evolution and gene regulation, and provide a bioengineering tool[6,19]. A model that accurately approximates the relationship between sequence and expression can serve as an 'oracle' in evolutionary studies to conduct and interpret in silico experiments[20–23], predict which regulatory mutations affect expression and fitness (when coupled with expression-to-fitness curves[11]), design or evolve new sequences with desired characteristics, determine how quickly selection achieves an expression optimum, identify signatures of selective pressures on extant regulatory sequences, visualize fitness landscapes and characterize mutational robustness and evolvability[2,4–6,24,25].

Here we address these long-standing problems by developing a framework for studying regulatory evolution and fitness landscapes (Fig. 1a) based on *Saccharomyces cerevisiae* promoter sequence-to-expression models.

Models predict expression from sequence

We begin by building models that predict gene expression given an 80-bp sequence of promoter DNA. To train these models, we measure the expression driven by promoter sequences using an approach we previously described[26], in which 80 bp of DNA are embedded within a promoter construct and the associated expression is assayed in the *S. cerevisiae* (Methods). We clone promoter sequences into an episomal low-copy-number yellow fluorescent protein (YFP) expression vector, transform them into yeast, culture the yeast in the desired medium, sort the yeast into 18 expression bins and sequence the promoters present from the yeast in each bin to estimate expression (Methods, Supplementary Information). To avoid biases[5] towards extant sequences, we measured the expression of 80-bp random DNA sequences, in which each base is randomly sampled from the four bases. For training data, we measured each of more than 30 million sequences in complex medium (yeast extract, peptone and dextrose (YPD); Methods) and more than 20 million sequences in defined medium (synthetic defined medium lacking uracil (SD-Ura)). Using the resulting pairs of sequences and measured levels of YFP expression, we trained convolutional neural network models ('convolutional models') that predict expression from sequence in each medium (Methods).

To show that the learned convolutional models generalize to new sequences, we predicted the expression for several sets of test sequences not seen during model training, and compared them to their experimentally measured levels (Methods). For these test sequences, we quantified expression in independent experiments using the same experimental approach and in the same medium. Our convolutional models had excellent prediction performance on native yeast promoter test sequences (Pearson's $r = 0.960$, $P < 5 \times 10^{-324}$, $n = 61{,}150$; Fig. 1b), and on multiple other test sets in both complex medium and defined medium (Extended Data Fig. 1).

These results represent a decrease in error of around 45% compared to the performance of biochemical models that we previously[26] trained on the same data (complex medium; native yeast promoter

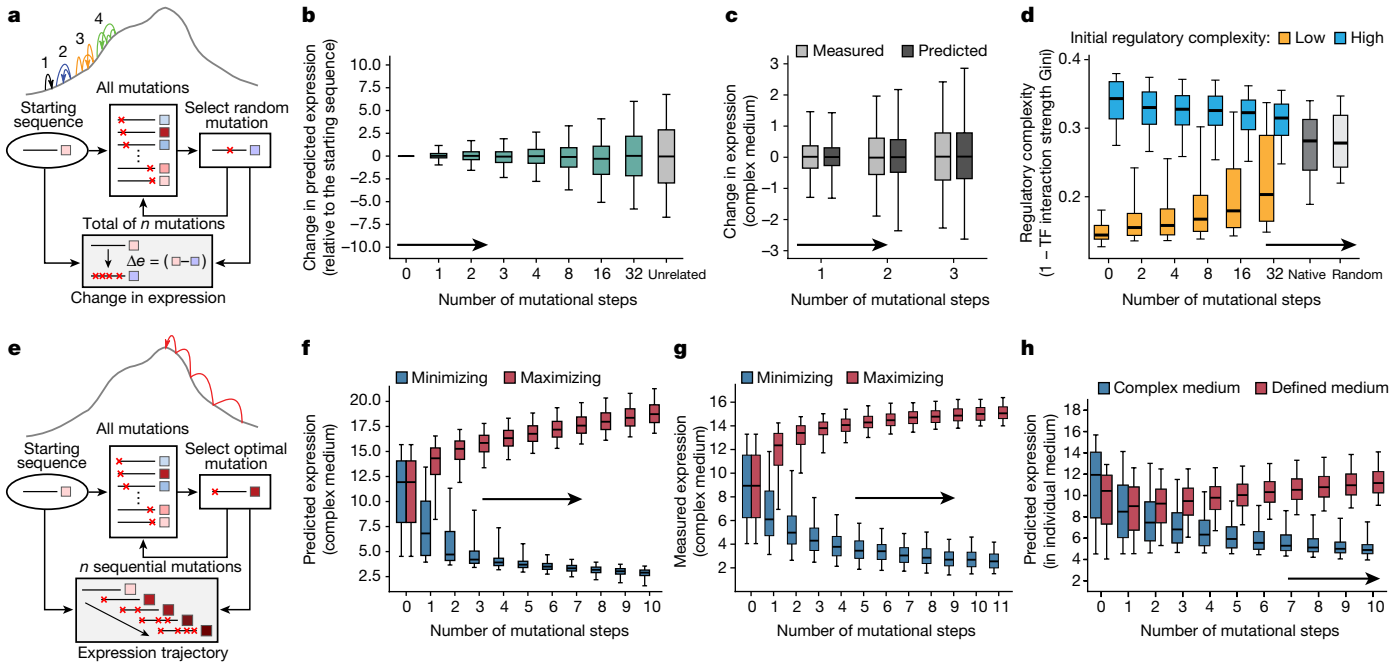


Fig. 2 | The evolutionary malleability of gene expression. a–**c**, Expression divergence under random genetic drift. **a**, Simulation procedure. **b**, Predicted expression divergence. Distribution of the change in predicted expression (*y* axis) for random starting sequences ($n = 5{,}720$) at each mutational step (*x* axis) for simulated trajectories. Silver bar: expression differences between unrelated sequences. **c**, Experimental validation. Distribution of measured (light grey) and predicted (dark grey) changes in expression in complex medium (*y* axis) for synthesized randomly designed sequences ($n = 2{,}983$) at each mutational step (*x* axis). **d**, Stabilizing selection on gene expression leads to moderation of regulatory complexity extremes. Regulatory complexity (*y* axis) of sequences from sequential mutational steps (*x* axis) under stabilizing selection to maintain the starting expression levels, in which the regulatory interactions of starting sequences are complex (blue; $n = 192$) or simple (orange, $n = 172$). Right bars: regulatory complexity for native (dark grey) and random (light grey) sequences. **e**–**g**, Sequences under SSWM regimes can rapidly evolve to expression optima. **e**, Simulation procedure. **f**, Predicted expression evolution. Distribution of predicted expression levels (*y* axis) in complex medium at each mutational step (*x* axis) for trajectories favouring high (red) or low (blue) expression, starting with native promoter sequences ($n = 5{,}720$). **g**, Experimental validation. Measured expression distribution in complex medium (*y* axis) for the synthesized sequences ($n = 10{,}322$ sequences; 877 trajectories) at each mutational step (*x* axis), favouring high (red) or low (blue) expression. Axis scales differ owing to variation in measurement procedure (Supplementary Information). **h**, Competing expression objectives constrain expression adaptation. Distribution of predicted expression (*y* axis) in complex (blue) and defined (red) medium at each mutational step (*x* axis) for a starting set of native promoter sequences ($n = 5{,}720$), optimizing for high expression in defined (red) and simultaneous low expression in complex (blue) medium. In **b**–**d**, **f**–**h**, midline, median; boxes, interquartile range; whiskers, 5th and 95th percentile range.

test sequences; Supplementary Notes, Methods). Other published genomic model architectures adapted to and trained using our data also had excellent performance (Supplementary Fig. 4a), highlighting the predictive power of deep neural network models trained using our large-scale data. Finally, the expression measurements were highly correlated for the same sequences between the two media (Pearson's $r = 0.978$; Extended Data Fig. 2a), and models trained on defined medium predicted expression in complex medium well (Pearson's $r = 0.966$; Extended Data Fig. 2b). However, for some sequences we expect differences between growth conditions (see below).

Models enable expression engineering

We leveraged the high predictive performance of our convolutional models for a synthetic biology application of gene expression engineering, by using model predictions as a 'fitness function' for genetic algorithms to design sequences with extreme expression values. We initialized the genetic algorithm with a population of 100,000 randomly generated samples from the sequence space, and simulated 10 generations to maximize (or minimize) the expression output from the convolutional model (Methods). We then synthesized the 500 sequences with the top predicted maximum (or minimum) expression levels and tested them experimentally. The genetic-algorithm-designed sequences drove, on average, more extreme expression than more than 99% of native sequences (99.6% for high expressing; 99.3% for low), with around 20% of designed sequences yielding more extreme expression than any native sequence tested (23.5% for high; 18.4% for low) (Fig. 1c). Thus, our sequence-to-expression model can be used to design sequences for gene expression engineering.

Expression diverges under genetic drift

We next assessed the evolutionary malleability of expression under different evolutionary scenarios: random genetic drift, stabilizing selection, and directional selection for extreme expression levels (Fig. 2). In each case, we first simulated the scenario, using our convolutional model to predict the expression for each sequence, and then tested the evolved sequences of the model experimentally, where possible (Methods).

We first simulated random genetic drift of regulatory sequences, with no selection on expression levels. We randomly introduced a single mutation in each random starting sequence, repeated this process for multiple consecutive generations and used our convolutional model to predict the difference in expression between the mutated sequences in each trajectory relative to the corresponding starting sequence (Fig. 2a–c). Expression levels diverged as the number of mutations increased, with 32 mutations in the 80-bp region resulting in nearly as different expression from the original sequence as two unrelated sequences (Fig. 2b). We validated our results experimentally by synthesizing sequences with zero to three random mutations and

measuring their expression in our assay (Methods). The experimental measurements closely matched our predictions in both complex medium (Fig. 2c) and defined medium (Extended Data Fig. 1e), both in expression change (Pearson's $r = 0.869$ and 0.847, respectively; Extended Data Fig. 1h, i) and expression level (Pearson's $r = 0.973$ and 0.963, respectively; Extended Data Fig. 1l, m).

Stabilizing selection tempers complexity

Although gene regulatory networks often appear to be highly interconnected[26,27], the sources of this regulatory complexity and how it changes with the turnover of regulatory mechanisms[28] remain unclear. We used our model to study the evolution of regulatory complexity in the context of stabilizing selection, which favours the maintenance of existing expression levels. We first quantified regulatory complexity, defined as 1 minus the Gini coefficient (a measure of inequality of continuous values within a population) of transcription factor regulatory interaction strengths. For this, we used an interpretable biochemical model that we previously developed[26] (Methods) because it has parameters that explicitly correspond to transcription factors, and we can directly query their contributions to model predictions. Next, starting with native sequences whose regulatory complexity is either extremely high (many transcription factors with similar contributions to expression) or low (few transcription factors contribute disproportionately to expression), and spanning a range of expression levels, we introduced single mutations into each starting native sequence for each of 32 consecutive generations, identified the sequences that conserved the original expression level using the convolutional model and selected one of them at random for the next generation. We then assessed the regulatory complexity of the evolved sequences.

As random mutations accumulated, the regulatory complexity of sequences starting at both complexity extremes shifted towards moderate complexities, closer to the averages for both random sequences and native sequences (Fig. 2d). This suggests that stabilizing selection on expression leads to a moderation of regulatory complexity, resulting from gradual drift in the roles of the different regulators, such as an increase in complexity due to a decrease in the relative contribution of one predominant transcription factor (for example, Abf1p for *AIF1*), or a decrease in complexity through smaller changes in a much larger number of sites (for example, *YDR476C*; Supplementary Fig. 8). The overall distribution of regulatory complexity of native yeast promoters is similar to that of random sequences (Fig. 2d), suggesting that there is little selection on the regulatory complexity of native sequences in a single environment.

Strong selection rapidly finds extremes

To study the effect of directional selection on expression, we simulated the strong-selection weak-mutation (SSWM) regime[29] (Fig. 2e, Methods), in which each mutation is either beneficial or deleterious (strong selection, with mutations surviving drift and fixing in an asexual population), and mutation rates are low enough to only consider single-base substitutions during adaptive walks (weak mutation). Starting with a set of native promoter sequences, at each iteration (generation), for a given starting sequence of length L, we considered all of its $3L$ single-base mutational neighbours, used our convolutional model to predict their expression and took the sequence with the largest increase (or separately, decrease) in expression at each iteration (generation) as the starting sequence for the next generation (Fig. 2e, Methods).

Sequences that started with diverse initial expression levels rapidly evolved to high (or separately, low) expression, with the vast majority evolving close to saturating extreme levels within three to four mutations in both the complex medium (Fig. 2f) and the defined medium (Extended Data Fig. 1f). Sequences took diverse paths to evolve either high or low expression (Supplementary Fig. 7). We validated these trajectories experimentally for select series of sequences (Fig. 2g, Extended Data Fig. 1g), measuring the expression driven by synthesized sequences from several generations along simulated mutational trajectories for complex medium (10,322 sequences from 877 trajectories) and defined medium (6,304 sequences from 637 trajectories). We observed extreme expression within three to four mutational steps, with high agreement between measured and predicted expression change (Extended Data Fig. 1j, k; Pearson's $r = 0.977$ and 0.948, respectively) and expression levels (Extended Data Fig. 1n, o; Pearson's $r = 0.980$ and 0.963, respectively) along the trajectories in both complex medium and defined medium. Thus, the evolution of *cis*-regulatory sequences is rapid and subject to diminishing returns epistasis[30].

Opposing objectives constrain adaptation

In contrast to the rapid evolution towards expression extremes, we found that evolution to satisfy two opposing expression requirements (one in each growth medium) was more constrained. A concrete example is the expression of the *URA3* gene: organismal fitness increases with increased *URA3* expression in defined medium lacking uracil, because Ura3p is required for uracil biosynthesis, but fitness decreases with increased *URA3* expression in complex medium containing 5-FOA, owing to Ura3p-mediated conversion of 5-FOA to toxic 5-fluorouracil (Extended Data Fig. 2c). To study this regime[31], we started with a set of native promoter sequences (and separately, a set of random sequences) and used the convolutional model to simulate SSWM trajectories (Methods) that maximize the difference in expression between the two media (defined and complex). Although the difference in expression increased with each generation (Extended Data Fig. 2d, e), most sequences achieved neither the maximal nor the minimal expression in either condition after 10 generations (Fig. 2h, Extended Data Fig. 2f), for both native and random starting sequences. The evolved sequences became enriched for motifs for transcription factors that are involved in nutrient sensing and metabolism, compared to the starting sequences (Extended Data Fig. 2g), suggesting that the model is taking advantage of subtle differential activity of certain regulators between the two conditions to evolve condition specificity. Thus, although evolving a sequence to achieve a single expression optimum requires very few mutations, encoding multiple opposing objectives in the same sequence is more difficult, which limits expression adaptation.

Transformer models enable inference at scale

We next turned to the evolution and evolvability of regulatory sequences in extant strains and species. This required us to predict expression for billions of sequences and, although our convolutional model had excellent predictive power, our implementation was limited in its scalability and incompatible with the tensor processing units (TPUs) available to us for larger-scale computational tasks (Methods). To enable large-scale expression prediction, we developed 'transformer' models that used transformer encoders[32] with other building blocks to help implicitly capture known aspects of regulation[33] (Methods, Supplementary Fig. 12). The transformer models had around 20 times fewer parameters than the convolutional models (Methods, Supplementary Information), predicted expression as well as the convolutional models (Extended Data Fig. 3) and better captured the propensity for expression to plateau under SSWM (Supplementary Fig. 19). The convolutional and transformer models had highly correlated predictions in both media (Supplementary Fig. 4e–h; Pearson's $r = 0.967$–0.985), and yielded equivalent conclusions from the analyses of genetic drift, directional selection and conflicting objectives (Extended Data Fig. 3, Supplementary Figs. 17, 18).

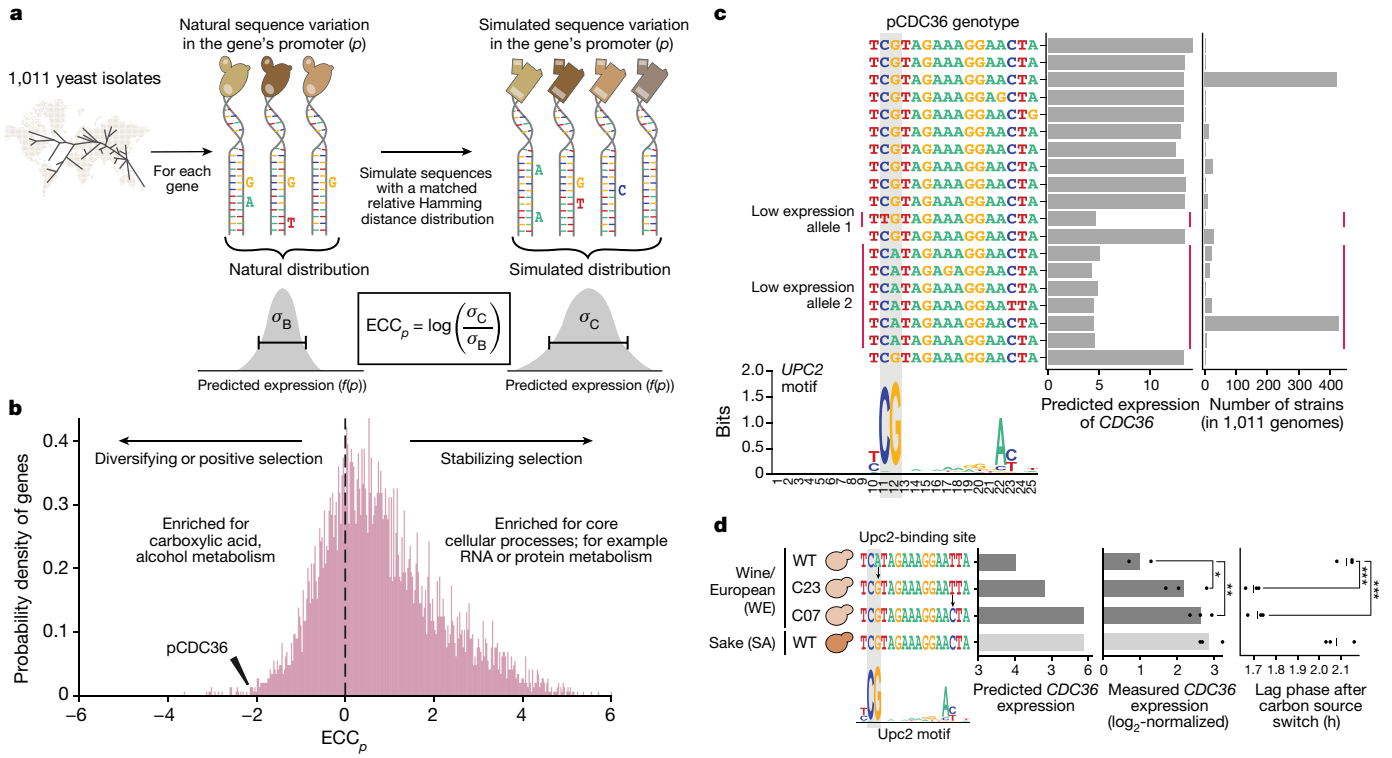


Fig. 3 | The ECC detects signatures of selection on gene expression using natural genetic variation in regulatory DNA. a, ECC calculation from 1,011 *S. cerevisiae* genomes[37]. '*p*' refers to a gene's promoter sequence; 'ECC_p' refers to the ECC for that gene. **b**, ECC distribution for *S. cerevisiae* genes. Frequency distribution of ECC values (*x* axis). Dashed line separates regions corresponding to disruptive or positive selection (left) and stabilizing selection (right). Gene Ontology (GO) terms enriched by the ECC ranking are shown. Arrowhead: ECC value for the *CDC36* promoter sequence. **c**, Convergent regulatory evolution in the *CDC36* promoter. Predicted expression (*x* axis, left bar plot) and associated number of strains (*x* axis, right bar plot) of all alleles among the analysed *CDC36* promoter sequence within 1,011 yeast isolates, along with an alignment of their Upc2p-binding site sequences (left; Upc2p-binding motif below). Red vertical lines: two independently evolved low-expressing alleles. Grey vertical boxes: key positions in the Upc2p motif with single nucleotide polymorphisms. **d**, Validation of *CDC36* promoter allele expression and organismal phenotype. Strains (*y* axis) with different Upc2p-binding site alleles for both model-predicted *CDC36* expression (left; predicted on the −170 to −90 region (relative to the TSS) to capture the entire Upc2p-binding site), measured *CDC36* expression (middle) and lag phase duration (right). WT, wild type. Points: biological replicates ($n = 3$); bars and vertical lines: means. Bar colour: strain background. Student's *t*-test *P* values, unpaired, equal variance, one-sided (expression, WE WT versus C23 $P = 0.044$, C07 $P = 6.69 \times 10^{-3}$) or two-sided (lag phase, WE WT versus C23 $P = 1.34 \times 10^{-4}$, C07 $P = 2 \times 10^{-4}$); $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

The expression conservation coefficient

We applied our sequence-to-expression transformer model to detect evidence of selective pressures on natural regulatory sequences, inspired by the way in which the ratio of non-synonymous ('non-neutral') to synonymous ('neutral') substitutions (d_N/d_S) in protein-coding sequences is used to estimate the strength and mode of natural selection[34]. By analogy[2,35], for regulatory sequences[2], we used the transformer model to quantitatively assess the effect of naturally occurring regulatory genetic variation on expression, compared to that expected with random mutations, and summarized this with an expression conservation coefficient (ECC) (Methods). To compute the ECC, we compared, for each gene's promoter, the standard deviation of the expression distribution predicted by the transformer model for a set of naturally varying orthologous promoters (σ_B) to the standard deviation of the expression distribution predicted for a matched set of random variation introduced to that promoter (σ_C; related to the mutational variance[36]; Fig. 3a). We define the ECC for a gene as $\log(\sigma_C/\sigma_B)$, such that a positive ECC indicates stabilizing selection on expression (lower variance in native sequences than expected by chance), a negative ECC indicates diversifying (disruptive) selection or local adaptation (greater variance in native sequences) and values near 0 suggest neutral drift.

We calculated the ECC for 5,569 *S. cerevisiae* genes using the natural variation observed across over 4.73 million orthologous promoter sequences from the 1,011 *S. cerevisiae* isolates[37] in the −160 to −80 region (relative to the transcription start site (TSS)), a critical location for transcription factor binding[38] and determinant of promoter activity[26] (Fig. 3a, b, Supplementary Table 1), using our transformer model to predict the expression for each sequence. To assess the robustness of the ECC values, we recomputed the ECC using multiple published sequence-to-expression model architectures that we adapted and trained using our data and found that models with similarly high predictive power resulted in similar ECC values (Supplementary Figs. 4b–d, 5g).

Over 70% of promoters had positive ECCs, suggesting stabilizing selection (and conserved expression) (binomial test $P < 10^{-215}$) (Fig. 3b), consistent with previous reports based on direct measurements of gene expression[39]. Genes with high ECCs were enriched in highly conserved core cellular processes (for example, RNA and protein metabolism) (Fig. 3b, Supplementary Table 2), and those with low ECCs were most enriched in processes related to carboxylic acid and alcohol metabolism (Fig. 3b, Supplementary Table 2), potentially reflecting adaptation of fermentation genes to the diverse environments of these isolates[37].

Discovering convergent evolution using the ECC

A notable example of predicted positive selection is the promoter of *CDC36* (*NOT2*; ECC = −2.138; Fig. 3b), which has common natural alleles with either low or high (predicted) expression across the isolates (Fig. 3c). Analysis of *CDC36* promoter sequences (Methods) suggests that low expression evolved at least twice independently,

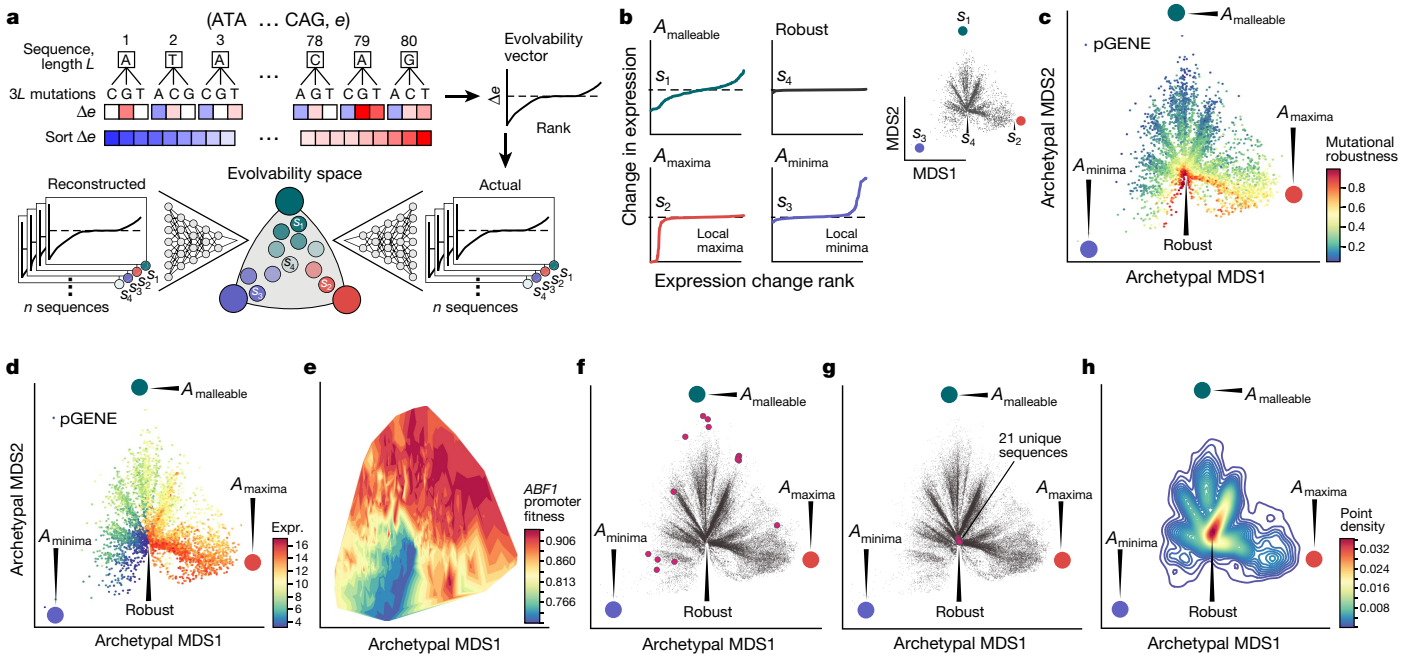


Fig. 4 | The evolvability vector captures fitness landscapes. **a**, Characterizing regulatory evolvability by computing an evolvability vector. Generating evolvability vectors for a sequence (top). Training an autoencoder with evolvability vectors to generate a 2D representation to visualize sequences in archetypal evolvability space (bottom). **b**, Evolvability archetypes discovered by the autoencoder. Left, evolvability vectors of the rank-ordered (*x* axis) predicted change in expression (*y* axis) for native sequences closest to each of the malleable (green), maxima (red) or minima (blue) archetypes and the 'robustness cleft' (black). Right, all native yeast (*S. cerevisiae* S288C) promoter sequences (grey points) projected onto the archetypal evolvability space by their evolvability vectors. Evolvability archetypes (coloured circles) and their closest native sequences (s_1–s_4 as on left) are marked. MDS, multidimensional scaling. **c**, **d**, Evolvability space captures mutational robustness and expression levels. Evolvability vectors (points) for the promoter sequences of each native yeast gene (pGENE) projected onto the evolvability space (archetypes are large coloured circles, as in **b**) and coloured by mutational robustness (**c**) or predicted expression levels (Expr.; **d**). **e**, *ABF1* promoter fitness landscape. Evolvability vectors of promoter sequences projected onto the evolvability space and coloured by computed fitness (Methods). **f**, **g**, Malleable promoter sequences dynamically traverse the evolvability space. Evolvability vector projections of native sequences (points) from all 1,011 *S. cerevisiae* isolates. Red points: natural promoter sequence variants for *DBP7*, the promoter closest to the malleable archetype (14 unique sequences; **f**) and for *UTH1*, the promoter closest to the robustness cleft (21 unique sequences; **g**). **h**, The robustness of native promoter sequences. Density (colour) of all native yeast promoter sequences when their evolvability vectors are projected onto the evolvability space.

resulting in two distinct variants with reduced expression (Fig. 3c, allele 1 and 2). Interrogation with the biochemical model[26] to identify factors that mediate these expression differences (Extended Data Fig. 4a) suggested that both low-expression alleles are explained by disruption of the same binding site for Upc2p, an ergosterol-sensing transcription factor (Fig. 3c). To validate this, we restored the putative Upc2p-binding site in a strain (Wine/European; WE) in which it is otherwise disrupted, then measured expression levels by quantitative PCR (qPCR), and measured growth after changing carbon source (Methods). Restoration of the Upc2p-binding site increased actual expression, confirming the prediction of the model (Pearson's $r = 0.96$, $P = 0.039$, $n = 4$; Fig. 3d). We hypothesized that these variants could alter the rate of transcriptional reprogramming when changing environments through Cdc36p-regulated mRNA turnover[40]. Indeed, restoration of the Upc2-binding site reduced the lag time of the strains to growth when switching carbon sources (Fig. 3d, Methods), and they grew to a higher culture density (Supplementary Fig. 10). Thus, convergent evolution of the *CDC36* promoter, discovered using the ECC, independently produced two alleles that result in similar perturbations to transcription factor binding, expression and growth.

ECC predicts expression conservation and fitness

ECC values were consistent with expression conservation as measured for yeast orthologues across clades at short (*Saccharomyces*), medium (Ascomycota) or long (mammals) evolutionary scales (Extended Data Fig. 4b). In *Saccharomyces*, 1:1 orthologues with conserved expression levels across species (as measured by RNA sequencing (RNA-seq))[41] had a significantly higher ECC (computed from the 1,011 yeast isolates) than genes whose RNA-seq expression levels were not conserved (two-sided Wilcoxon rank-sum $P = 3.1 \times 10^{-4}$; Extended Data Fig. 4b, Methods). Next, we performed RNA-seq across 11 Ascomycota yeast species (Methods), and found that 1:1 orthologues with conserved expression across Ascomycota had significantly higher ECC values (Extended Data Fig. 4b; $P = 1.16 \times 10^{-6}$). Finally, the 1:1 orthologues of genes that were conserved in expression across mammals also had high ECC values in the 1,011 *S. cerevisiae* isolates[42] (Extended Data Fig. 4b; $P = 1.07 \times 10^{-4}$, Methods). Thus, although all 1:1 yeast–mammal orthologues are likely to be critical to an organism's fitness, only a subset of these may be under stabilizing selection on expression, and this subset tends to be under such selection in both yeasts and mammals. Thus, the ECC quantifies stabilizing selection on expression in yeast and may predict stabilizing selection on the expression of orthologues in other species.

Genes with higher ECCs also had a stronger effect on fitness in *S. cerevisiae* when their expression level was changed. We interrogated the total variation of previously measured expression-to-fitness curves[11] to calculate a 'fitness responsivity' score that captures the dependence of fitness on expression of a gene(Extended Data Fig. 5, Methods). Fitness responsivity was significantly positively correlated with the ECC (Supplementary Fig. 2e; $P = 0.003$, Spearman's $\rho = 0.326$). Fitness

responsivity was not associated with regulatory sequence divergence per se across the promoter sequence (as estimated by mean Hamming distance among orthologous promoters; Methods, Supplementary Fig. 2d; $P = 0.46$, Spearman's $\rho = 0.083$). Thus, although stabilizing selection on gene expression (as captured by the ECC) can shape the types of mutations that accumulate in the population, it may have little effect on the overall rate at which mutations accumulate in promoter regions within populations, which has been previously used to test for evidence of selection.

Stabilizing selection shapes robustness

Although a gene's ECC (computed from the natural genetic variation in regulatory DNA) represents the imprint of its evolutionary history, its mutational robustness (assessed directly from the gene's promoter sequence) should describe how future mutations would affect its expression[43]. Across all native yeast promoters, the magnitude of expression changes predicted by the transformer model owing to single-base-pair mutations follows a power law with an exponent of 2.252 (standard error of fit $\sigma = \pm 0.002$, $P = 2.4 \times 10^{-263}$), such that a small number of mutations have an outsized effect on expression (around 10% of mutations account for around 50% of the changes in expression; Extended Data Fig. 4d). In individual genes, the distribution can vary substantially (see below).

For a given promoter sequence, we defined the mutational robustness of a sequence of length L as the percentage of its $3L$ single-nucleotide mutational neighbours predicted by the transformer model to result in a negligible change in expression (Extended Data Fig. 4c, Methods), following previous definitions of mutational robustness[25,43]. The mutational robustness of the promoter sequence of a gene was positively correlated with the gene's fitness responsivity (Supplementary Fig. 2f; Spearman's $\rho = 0.476$, $P = 8.18 \times 10^{-6}$), suggesting that fitness-responsive genes have evolved more mutationally robust regulatory sequences. Mutational robustness—which, unlike the ECC, is computed for single sequences without a set of variants across a population—was also correlated to the ECC (Supplementary Fig. 2g; Spearman's $\rho = 0.515$, $P = 9.99 \times 10^{-7}$). Similarly, the promoter sequences of yeast genes with conserved expression across *Saccharomyces* strains[41], Ascomycota species or mammals[42] had higher mutational robustness ($P = 8.4 \times 10^{-3}$, $P = 6.5 \times 10^{-5}$ or $P = 0.00377$, respectively, two-sided Wilcoxon rank-sum test).

Thus, genes with expression levels that are under stabilizing selection have regulatory sequences that tend to be more robust to the effects of mutations, which may reflect their history and constrain their future.

Fitness landscapes in evolvability space

Mutational robustness enables the exploration of new genotypes that could subsequently facilitate adaptation and thus promote evolvability—the ability of a system to generate heritable phenotypic variation[25]. To characterize regulatory evolvability, we extended our description of mutational robustness by representing each sequence using a sorted vector of expression changes (predicted by the transformer model) that are accessible through single-nucleotide mutations (Fig. 4a, Methods). This 'evolvability vector' captures the capacity for changes in genotype to alter expression phenotype, in line with previous definitions of evolvability[25].

We next asked whether regulatory evolvability vectors fell into distinct classes by identifying evolvability 'archetypes'. Archetypes[44] represent the extremes of canonical patterns, such that the evolvability vector of each individual sequence can be represented by its similarity to each of several archetypes representing these extremes. Applying this paradigm, we used our transformer model to compute evolvability vectors for a new random sample of a million sequences and then learned a two-dimensional representation of these evolvability vectors (referred to as the 'evolvability space') using an autoencoder[45] (Fig. 4a, Methods). This archetypal evolvability space, which is bounded by a simplex whose vertices represent evolvability archetypes (Fig. 4a, Methods) and in which the evolvability vector of each sequence is a single point, allows us to effectively visualize arbitrarily large sequence spaces in two dimensions.

Three archetypes captured most of the variation in evolvability vectors (Extended Data Fig. 6a, b, Methods), corresponding to local expression minimum (A_{minima}), local expression maximum (A_{maxima}) and malleable expression ($A_{\text{malleable}}$) (Fig. 4b). A_{minima} and A_{maxima} correspond to sequences in which most $3L$ mutational neighbours do not change expression, and the ones that do, increase it (for A_{minima}) or decrease it (for A_{maxima}). Conversely, for $A_{\text{malleable}}$ sequences, most $3L$ mutational neighbours change expression and are equally likely to decrease or increase it (Fig. 4b). In addition to these three archetypes, mutationally robust sequences were present as a central cleft in the evolvability space (Fig. 4b, c; 'robust'). The evolvability space also distinguishes native regulatory sequences by their associated expression level (Fig. 4d), with intermediate expression more likely to be near the malleable archetype ($A_{\text{malleable}}$) and depleted near the robustness cleft (Fig. 4d, Supplementary Information).

The location of sequences in evolvability space reflects the selective pressures that operate on the sequence. Sequences under strong stabilizing selection on gene expression tend to be located far away from the malleable archetype: there is a strong negative correlation between malleable archetype proximity and mutational robustness (Extended Data Fig. 6c, e; Spearman's $\rho = -0.746$, $P = 1.97 \times 10^{-15}$), the ECC (Extended Data Fig. 6d, f, g; $\rho = -0.596$, $P = 5.4 \times 10^{-9}$), fitness responsivity (Extended Data Fig. 6h; $\rho = -0.413$, $P = 1.4 \times 10^{-4}$) and expression conservation across species as measured by RNA-seq (*Saccharomyces*, $P = 0.000251$; Ascomycota, $P = 0.00002$; mammals, $P = 0.00114$; two-sided Wilcoxon rank-sum test).

To visualize promoter fitness landscapes in two dimensions we combined our sequence-to-expression transformer model with previously measured expression-to-fitness curves[11], and integrated them with the two-dimensional archetypal evolvability space (Fig. 4e, Extended Data Fig. 7, Methods). Unlike previous visualizations of fitness landscapes, which group sequences by their sequence similarity, here, sequences are arranged by the similarity in their evolvability. This approach effectively visualizes arbitrarily large sequence spaces in two dimensions, and groups sequences by their evolutionary properties. This addresses the challenges otherwise posed by sequence similarity-based landscapes, as highly similar regulatory sequences can have different functional properties (for example, owing to a loss of a transcription factor binding site), whereas very different sequences can be functionally similar (for example, owing to shared transcription factor binding sites). When organismal fitness is available for a particular gene and overlaid on the landscape (Fig. 4e, Extended Data Fig. 7), the resulting patterns depend on both the condition-specific sequence-to-expression function (for example, governing colour (fitness) through predicted expression, and embedded position, through evolvability) and the gene- and condition-specific expression-to-fitness functions.

Finally, we studied how natural yeast sequences explored evolutionary space, by placing the evolvability vectors of each of set of orthologous promoters of the 1,011 sequenced *S. cerevisiae* isolates[37] in the archetypal evolvability space. When a gene's promoter from one strain is near the malleable archetype, its orthologues in the other strains tend to broadly distribute in the evolvability space (Extended Data Fig. 6i), but avoid the robustness cleft (for example, the *DBP7* promoter from strain S288C; Fig. 4f). Conversely, when a promoter is near the robustness cleft (for example, the *UTH1* promoter from S288C), so are its orthologues (Fig. 4g, Extended Data Fig. 6i). Using in silico mutagenesis to interpret our model, we found that the *DBP7* promoter is particularly malleable partly as a result of an intermediate affinity Rap1p-binding site, whereby the mutations with the strongest effect increased or decreased the Rap1p affinity for this site, thus affecting expression (Extended Data Fig. 8a).

By contrast, the *UTH1* promoter requires many sequential mutations, each of which has a minimal effect individually, to reduce expression appreciably (Extended Data Fig. 8b). This could reflect the ways in which stabilizing selection constrains evolvability: promoters that are not under strong stabilizing selection explore expression space more freely and can quickly adapt to a new expression optimum, as the population is likely to already contain multiple alleles that achieve diverse expression levels (Fig. 4f). Notably, many of the native sequences in *S. cerevisiae* are near the robustness cleft (Fig. 4h).

Thus, the evolvability vector, which can be computed using our model directly for any sequence (without any population genetics data), encodes information about the evolutionary history of a sequence and its potential futures.

Discussion

Here we have presented a framework that addresses fundamental questions in the evolution and evolvability of regulatory sequences[2,25]. Our models, which were developed using a combination of large-scale random-sequence libraries, sensitive reporter assays and deep learning (Methods), are useful as 'oracles' for model-guided biological sequence design[19], and for answering key questions in the study of fitness landscapes[4–6], evolutionary malleability of expression and its variation across strains and species[2], mutational robustness[43] and evolvability[25]. The framework presented here will facilitate advances in synthetic biology, cell and gene therapy and metabolic engineering in addition to the study of evolution.

Previous studies suggested that evolution favours more complex regulatory solutions[46], but we have shown that if stabilizing selection acts only on expression, regulatory complexity extremes gradually move towards the moderate complexity levels that are observed in native and random sequences (Fig. 2d). This supports a model in which most extant regulatory sequences evolved by sampling constraint-satisfying solutions in proportion to their frequency in the sequence space, without specific consideration of the complexity of the solution.

In our study, evolving condition specificity in a promoter sequence was much slower than simply modifying the expression level. Some yeast genes achieve condition specificity by including multiple binding sites for condition-responsive transcription factors. For example, the *GAL1-10* upstream activating sequence contains multiple binding sites for the galactose-responsive Gal4; these are conserved across millions of years, which suggests an ancient origin[47]. Because the size of the regulatory region restricts the number of transcription factor binding site locations, including more transcription factors and more regulatory sequences per gene (for example, enhancers) may be required for more complex regulatory programs that are observed in higher eukaryotes[48].

The d_N/d_S ratio has been used extensively to characterize the evolutionary rates of protein-coding genes[34], and we developed an analogous[2,35] coefficient—the ECC—for detecting evidence of selection on expression from natural variation across multiple orthologous regulatory sequences in strains of one species. The ECC complements and extends existing measures of expression conservation because it integrates across the regulatory sequence and is not limited to specific transcription factors or binding motifs, does not require additional experiments to test the functions of mutations for each regulatory region and does not rely on detecting non-uniformity in mutation distributions.

Complementing the ECC, mutational robustness as calculated with our model is predictive of selective pressures on individual sequences (Supplementary Fig. 2f, g). Although we find that strong constraint on the function of regulatory sequences can shape them to be robust to future mutations, it is unlikely that robustness itself is the selected trait, as increased robustness to future mutations is likely to be of little marginal benefit[43]. Instead, this may reflect a secondary benefit of having evolved decreased expression noise[49,50], or another as-yet-unknown mechanism. It may also reflect the fact that the sequences of some ancestral promoters may be similar to the mutational neighbours of extant sequences, and, if selective constraints on expression have remained stable, these ancestral and extant sequences are likely to have similar expression levels.

On the basis of our model-derived evolvability vectors, sequences spanned an evolvability spectrum from robust to malleable (Fig. 4c, d, f–h), and for native regulatory sequences, the magnitudes of accessible mutation effects follows a power law. Evolvability vectors also enable visualizations of fitness landscapes[4] (Fig. 4e, Extended Data Fig. 7), and future work can further improve our understanding of their topography[4,5].

Our sequence-to-expression models are at present limited by regulatory region and species. For instance, sequence mutations that affect other regulatory mechanisms (for example, genomic context, mRNA processing and degradation, regulation by RNA-binding proteins and translational efficiency) can compensate for those that affect transcription. Although our models emulated the biological process of our experimental system, as demonstrated by their predictive power, future interpretability studies will shed further light on molecular mechanisms. Finally, for multicellular organisms, selection acts simultaneously on expression levels in many different cell types and environments. As models of gene regulation are created for other species, environments and regulatory regions, our framework will lead to further insights into regulatory evolution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04506-6.

1. Wittkopp, P. J. & Kalay, G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2011).
2. Hill, M. S., Vande Zande, P. & Wittkopp, P. J. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* **22**, 203–215 (2021).
3. Fuqua, T. et al. Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587**, 235–239 (2020).
4. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
5. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
6. de Visser, J. A. G. M., Elena, S. F., Fragata, I. & Matuszewski, S. The utility of fitness landscapes and big data for predicting evolution. *Heredity* **121**, 401–405 (2018).
7. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66–74 (2010).
8. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6**, 119–127 (2005).
9. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
10. Venkataram, S. et al. Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast. *Cell* **166**, 1585–1596 (2016).
11. Keren, L. et al. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* **166**, 1282–1294 (2016).
12. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
13. Ogden, P. J., Kelsic, E. D., Sinai, S. & Church, G. M. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **366**, 1139–1143 (2019).
14. Pitt, J. N. & Ferré-D'Amaré, A. R. Rapid construction of empirical RNA fitness landscapes. *Science* **330**, 376–379 (2010).
15. Shultzaberger, R. K., Malashock, D. S., Kirsch, J. F. & Eisen, M. B. The fitness landscapes of *cis*-acting binding sites in different promoter and environmental contexts. *PLoS Genet.* **6**, e1001042 (2010).
16. Mustonen, V., Kinney, J., Callan, C. G. & Lässig, M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl Acad. Sci. USA* **105**, 12376–12381 (2008).
17. Hartl, D. L. What can we learn from fitness landscapes? *Curr. Opin. Microbiol.* **0**, 51–57 (2014).
18. Otwinowski, J. & Nemenman, I. Genotype to phenotype mapping and the fitness landscape of the E. coli lac promoter. *PLoS ONE* **8**, e61570 (2013).

19. Sinai, S. & Kelsic, E. D. A primer on model-guided exploration of fitness landscapes for biological sequence design. Preprint at https://arxiv.org/abs/2010.10614 (2020).
20. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
21. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
22. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *Proc. 34th International Conference on Machine Learning* 3145–3153 (2017).
23. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
24. Fragata, I., Blanckaert, A., Louro, M. A. D., Liberles, D. A. & Bank, C. Evolution in the light of fitness landscape theory. *Trends Ecol. Evol.* **34**, 69–82 (2019).
25. Payne, J. L. & Wagner, A. The causes of evolvability and their evolution. *Nat. Rev. Genet.* **20**, 24–38 (2019).
26. de Boer, C. G. et al. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).
27. Crocker, J. et al. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191–203 (2015).
28. Habib, N., Wapinski, I., Margalit, H., Regev, A. & Friedman, N. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. *Mol. Syst. Biol.* **8**, 619 (2012).
29. Gillespie, J. H. Molecular evolution over the mutational landscape. *Evolution* **38**, 1116–1129 (1984).
30. Jerison, E. R. & Desai, M. M. Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. *Curr. Opin. Genet. Dev.* **35**, 33–39 (2015).
31. Sæther, B.-E. & Engen, S. The concept of fitness in fluctuating environments. *Trends Ecol. Evol.* **30**, 273–281 (2015).
32. Vaswani, A. et al. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 5998–6008 (Curran Associates, 2017).
33. Weirauch, M. T. et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134 (2013).
34. Yang, N. & Bielawski, N. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
35. Moses, A. M. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. *BMC Evol. Biol.* **9**, 286 (2009).
36. Rifkin, S. A., Houle, D., Kim, J. & White, K. P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* **438**, 220–223 (2005).
37. Peter, J. et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**, 339–344 (2018).
38. Erb, I. & van Nimwegen, E. Transcription factor binding site positioning in yeast: proximal promoter motifs characterize TATA-less promoters. *PLoS One* **6**, e24279 (2011).
39. Gilad, Y., Oshlack, A. & Rifkin, S. A. Natural selection on gene expression. *Trends Genet.* **22**, 456–461 (2006).
40. Alhusaini, N. & Coller, J. The deadenylase components Not2p, Not3p, and Not5p promote mRNA decapping. *RNA* **22**, 709–721 (2016).
41. Yang, J.-R., Maclean, C. J., Park, C., Zhao, H. & Zhang, J. Intra and interspecific variations of gene expression levels in yeast are largely neutral: (Nei Lecture, SMBE 2016, Gold Coast). *Mol. Biol. Evol.* **34**, 2125–2139 (2017).
42. Chen, J. et al. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.* **29**, 53–63 (2019).
43. Payne, J. L. & Wagner, A. Mechanisms of mutational robustness in transcriptional regulation. *Front. Genet.* **6**, 322 (2015).
44. Shoval, O. et al. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336**, 1157–1160 (2012).
45. van Dijk, D. et al. Finding archetypal spaces using neural networks. *IEEE International Conference on Big Data 2634-2643* (2019).
46. He, X., Duque, T. S. P. C. & Sinha, S. Evolutionary origins of transcription factor binding site clusters. *Mol. Biol. Evol.* **29**, 1059–1070 (2012).
47. Cliften, P. F. et al. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11**, 1175–1186 (2001).
48. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
49. Lehner, B. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol. Syst. Biol.* **4**, 170 (2008).
50. Metzger, B. P. H., Yuan, D. C., Gruber, J. D., Duveau, F. & Wittkopp, P. J. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521**, 344–347 (2015).

Methods

Experimental measurement of sequence–expression pairs using a Sort-seq strategy

We experimentally measured expression using a Sort-seq massively parallel reporter assay[2,3,51–59] strategy called the gigantic parallel reporter assay (GPRA) that we previously described[26] (Supplementary Fig. 1). In brief, for each set of expression measurements mentioned, random or designed single-stranded oligonucleotides were ordered from IDT (random; Supplementary Table 3) or Twist Biosciences (designed; sequences on the Gene Expression Omnibus (GEO); accession GSE163045), cloned into the promoter of a YFP gene within a CEN plasmid (Addgene: 127546) as previously described[26] and transformed into yeast (strain Y8205 for the training dataset of random sequences, and strain S288C::*ura3* for all the rest of the sequences measured; the full list of yeast strains used is available in Supplementary Table 4). The library is maintained in yeast as an episomal low-copy-number plasmid. It was previously reported that the expression measurements are highly correlated with expression levels as measured using integrated reporters ($R^2 = 0.97$)[54]. Yeast were grown in continuous log phase, diluting as necessary to maintain an optical density (OD) between 0.05 and 0.6 for 8–10 generations up until the time of collection. Cells were collected, washed once in ice-cold PBS, and kept on ice in PBS until sorting. Cells were sorted into 18 uniformly sized expression bins covering the majority of the expression distribution. After sorting, cells were regrown in SD-Ura until saturation, plasmids were isolated and the sequencing libraries created were sequenced with a 150 cycle NextSeq kit. For libraries with random 80-bp sequences, sequences were consolidated as previously described[26]. Reads from other (defined, non-random) libraries were aligned to the pre-defined sequences using Bowtie2[60], including only read pairs that perfectly matched a designed sequence. For each sequence, the expression level was the average of the expression bins in which it was observed, weighted by the number of times it was observed in each bin. These expression measurements were carried out separately in defined medium lacking uracil (SD-Ura (Sunrise Science, 1703-500)) and complex medium (YPD: yeast extract, peptone, dextrose).

Architecture of the convolutional model

A deep neural network model[20,21,23,61–69] with convolutional layers was constructed and used for designing sequences with high and low expression (Fig. 1c), and running evolutionary simulations under stabilizing selection, genetic drift and SSWM (Fig. 2) for each condition. These designed sequences, the expression of which was experimentally quantified (for example, Figs. 1c, 2d, g), were designed using models with the following architecture.

Input The input is the DNA sequence (*s*) represented in one-hot encoding. Input shape: (1, 110, 4).

Convolution block For the forward and reverse strand, separately,
- Strand-specific convolution layer 1. Kernel shape: (1, 30, 4, 256)
- Strand-specific convolution layer 2. Kernel shape: (30, 1, 256, 256)
- Concatenation of features from the forward and reverse strand
- Convolution layer 3. Kernel shape: (30, 1, 512, 256)
- Convolution layer 4. Kernel Shape: (30, 1, 256, 256)
- A bias term and an rectified linear unit (ReLU) activation was added to each convolution layer in this block.

Fully connected layers Fully connected layer 1. Kernel Shape: (110*256, 256).
- Fully connected layer 2. Kernel Shape: (256, 256)
- A bias term and a ReLU activation were added to each layer in this block.

Output Linear combination of the 256 features extracted as a result of all the previous operations on the sequence (*s*) to generate the predicted expression (*e*).

Every fully connected layer was *L2* regularized with a 0.0001 weight and had a dropout probability of 0.2.

Training of the convolutional model

For training, 20,616,659 random sequences for the defined medium and 30,722,376 random sequences for the complex medium (each to train a separate model) were used, along with their experimentally measured expression as described above. A mini-batch size of 1,024 was used for training and a mean squared error loss was optimized using the Adam optimizer[70] with an initial learning rate of 0.0005. The model was trained for five epochs. Model architecture was written in TensorFlow[71] 1.14 using Python 3.6.7. The convolutional model used TensorFlow graphs and sessions in its implementation and was thus incompatible with the TPUs[72]. These convolutional models (for both media) were used for all the predictions in Figs. 1, 2, Extended Data Figs. 1, 2.

The models were tested by predicting expression on sequences that the model had never seen before (Supplementary Fig. 21) that were measured in separate experiments, in which the library was lower complexity (fewer sequences) than the experiments that generated the training data, such that the expression associated with each sequence was measured with high accuracy (around 100 yeast cells per sequence on average). The test libraries included random, native (that is, present in the yeast genome) and designed sequences.

Training and evaluation were carried out on four Tesla M60 GPUs. All code for training and using the convolutional model is available here: https://github.com/1edv/evolution/tree/master/manuscript_code/model/gpu_only_model.

Architecture of the transformer model

A transformer model[23,32,73] was developed to run inference faster than the convolutional model, as needed for the evolutionary analyses in Figs. 3, 4. The transformer model had around 20 times fewer parameters (around 1.3 million, compared to the around 24 million parameters of the convolutional model) and was able to leverage TPUs for computation. Transformer models are used in all the analyses in Figs. 3, 4, Extended Data Figs. 3, 4, 6–8. Benchmarking analyses and ablation analyses for the transformer model are available in the Supplementary Information.

The deep transformer model has the following architecture (Supplementary Fig. 12).

Input The input is the DNA sequence (*s*) represented in one-hot encoding. Input shape: (110, 4).

Convolution block The convolution block is constructed in the following order (Supplementary Fig. 12b):
- Revere complement aware 1D convolution to get features for the forward and reverse strands. Kernel shape: (30, 4, 256).
- Batch normalization
- ReLU
- Concatenation of features from the forward and reverse strand
- 2D convolution: convolve over the combined features from both the strands to capture interactions between strands. Kernel shape: (2, 30, 4, 256)
- Batch normalization
- ReLU
- 1D convolution. Kernel shape: (30, 64, 64)
- Batch normalization
- ReLU

Transformer encoder blocks Two transformer encoder blocks[32,74,75] are constructed in the following order (Supplementary Fig. 12c):

- Multi-head attention: eight heads, capturing relations between features from different positions of (*s*) to compute a representation for the features extracted from the convolution block from (*s*).
- Residual connection
- Layer normalization
- Feed forward layer with eight units
- Residual connection
- Layer normalization

Bidirectional LSTM layer A bidirectional long short-term memory (LSTM) layer to capture the long-range interactions between different regions of the sequence with eight units and 0.05 dropout probability.

Fully connected layers Two fully connected layers with 64 hidden units, each consisting of ReLU and dropout (0.05 dropout probability). See Supplementary Fig. 12d.

Output Linear combination of 64 features extracted as a result of all the previous operations on the sequence (*s*) to generate the predicted expression (*e*).

Training of the transformer model

A total of 20,616,659 random sequences (defined medium) and 30,722,376 random sequences (complex medium), along with their experimentally measured expression, were used to train separate models for each medium. Model architecture was written in TensorFlow[71] 1.14 using Python 3.6.7 with multiple open source libraries (citations, where relevant, are included in the code for them). A mini-batch size of 1,024 was used for training and a mean squared error loss was optimized using a RMSProp optimizer[76] with a learning rate of 0.001. The stopping criterion monitored was the '*r*-squared' value and the model was allowed to train for 10 epochs without improvement before stopping training. Training was carried out on a Google Cloud Tensor Processing Unit (TPU)[72] v.3-8. Evaluation was carried out on four Tesla M60 GPUs. The model architecture visualization was generated using Netron 4.5.1. All processed data and models are publicly available on Zenodo at https://zenodo.org/record/4436477 and all code is available on GitHub at https://github.com/1edv/evolution/tree/master/manuscript_code/model/tpu_model. Transformer models are used in all the analyses in Figs. 3, 4, Extended Data Figs. 3, 4 6–8.

The models were tested by predicting expression on test sequences that the model had never seen before (Supplementary Fig. 21) that were measured in separate experiments, which included random, native (that is, present in the yeast genome) and designed sequences. To obtain expression measurements for each tested sequence that are more accurate than those from the high-complexity training data experiment, library complexity was limited such that each test promoter sequence is observed in around 100 yeast cells (Methods, Supplementary Information).

Gene expression engineering using a genetic algorithm for sequence design

To design[77–80] new sequences with desired expression, a genetic algorithm was implemented with the distributed evolutionary algorithms in Python (DEAP package)[81]. The mutation probability and the two-point crossover probability were set to 0.1 and the selection tournament size was 3. The initial population size was 100,000 and the genetic algorithm was run for 10 generations. The convolutional model was used as the basis for the objective function for the genetic algorithm, which was maximized for high expression and minimized for low expression (maximizing negative predicted expression). The top 500 sequences were synthesized (by Twist Biosciences) and expression was measured experimentally using our reporter assay, as described above.

Characterizing random genetic drift

Simulation of random genetic drift (Fig. 2a) was initialized with a set of 5,720 random sequences, in generation 0. For each sequence in this starting set, a new single sequence was randomly picked from its 3*L* mutational neighbourhood (the set of all sequences at a Hamming distance of 1 from a sequence of length *L*) and the difference in expression between the new sequence and the starting sequence was calculated using the convolutional model (Fig. 2b). This was done for each starting sequence to get generation 1. Each subsequent generation *n* was produced by picking a single sequence randomly from the 3*L* mutational neighbourhood of each sequence in the preceding generation $n-1$. The simulation was carried out for 40 generations. Simulations were also subsequently repeated with the transformer model (Extended Data Fig. 3f), yielding concordant results.

For experimental validation, 1,000 random starting sequences were synthesized, introducing between 1 and 3 random mutations to these sequences. The expression levels of starting and mutated sequences were measured in both complex medium and defined medium experimentally using our reporter assay. For 990 of these 1,000 starting sequences, experimental measurements were available for all 3 mutational distances. In addition, 20 (median) separate single mutations were introduced to each of 196 native sequences, the sequences were synthesized and their associated expression was measured similarly for both media; these were also included in the boxes for one mutational step in Fig. 2c, Extended Data Fig. 1e.

Characterizing the regulatory complexity of a sequence

To estimate the regulatory complexity[82,83] of a sequence, the Gini coefficient of the regulatory interaction strengths for each transcription factor was calculated. A new biochemical model was first trained with our defined medium data to complement the existing one trained on complex medium, using our published model architecture of transcription factor binding and position-aware activity[26] and the training procedure previously described[26] (Supplementary Notes). The regulatory interaction strength was then individually calculated for each regulator by setting the concentration parameter for that transcription factor (individually) to 0 in the learned model, and the biochemical model was used to quantify the resulting change in expression, as previously described[26]. The resulting vector of interaction strengths was used to calculate a Gini coefficient for each sequence, separately for the complex and defined medium models. The Gini coefficient is a measure of inequality of continuous values within a population, most commonly applied to wealth or income, and ranges from 0 (all members of the population have equal wealth) to 1 (the wealth of a population is held by a single individual). Regulatory complexity for a sequence is then 1 − Gini, such that 1 indicates that all transcription factors contribute equally to the regulation of the gene and 0 indicates that a single transcription factor is solely responsible for its regulation. As starting points for our trajectories, 200 native promoter sequences (from −160 to −80, relative to the TSS) were chosen with relatively high regulatory complexity and another 200 were chosen with relatively low regulatory complexity, spanning the range of predicted expression levels, as starting points for our trajectories.

Trajectories for stabilizing selection on gene expression were designed using the convolutional model (Fig. 2d). Here, all sequences were required to maintain a predicted expression level within 0.5 of the original expression levels at all steps along the trajectory. There was no explicit constraint on regulatory complexity in this simulation of stabilizing selection. To ensure that expression was unchanged,

expression levels were measured experimentally for sequences along a trajectory at growing mutational steps from the initial sequence (2, 4, 8, 16, 32 mutations). Any trajectories for which an expression measurement was missing for any experimentally tested sequence were excluded from all analyses, retaining 172 trajectories with initial low regulatory complexity and 192 trajectories with initial high regulatory complexity. Testing whether observed trends in regulatory complexity were affected by the degree to which expression was either predicted (by the transformer model for 1–32 mutations) or observed (by the experiment at 2, 4, 8, 16 or 32 mutations) to be conserved, showed that the trends were robust to the degree of expression conservation (Supplementary Fig. 11).

Characterizing directional trajectories under SSWM

Simulations of trajectories under an SSWM[84–86] regime were initialized with a set of native yeast promoter sequences (defined here as the subset from −160 to −80 relative to the TSS for all the genes in the yeast reference genome for which we had a good TSS estimate (see supplementary table 3 in ref. [26]) as the starting generation 0. For each sequence in generation n, the sequence from its 3L mutational neighbourhood that had the maximal (or separately, minimal) predicted expression using the convolutional model was picked to get generation $n+1$. The simulation was carried out for 10 rounds separately in the complex medium (Fig. 2f) and in the defined medium (Extended Data Fig. 1f). The simulations were subsequently repeated using the transformer model (Extended Data Fig. 3i, j).

For experimental validation, a subset of sequences from several generations was synthesized along mutational trajectories simulated by the convolutional model for complex medium (10,322 sequences from 877 trajectories, 805 of which had every sequence along the trajectory successfully measured) and one for defined medium (6,304 sequences from 637 trajectories, 591 of which had every sequence along the trajectory successfully measured) and their expression was measured in the corresponding medium experimentally using our reporter assay (Fig. 2g, Extended Data Fig. 1g).

Measuring the *URA3* expression-to-fitness relationship

Two complementary environments were studied with opposite selective pressures on the expression of *URA3* (encoding an enzyme responsible for uracil synthesis): defined medium, in which organismal fitness increases with gene expression (up to saturation); and complex medium + 5-FOA, in which fitness decreases with Ura3p expression.

Convolutional models trained on defined and complex media were used to choose a set of 11 sequences that span a broad range of predicted expression levels in the two media when cloned into a YFP expression vector[26]. The relationship between expression of *URA3* and organismal fitness in yeast was estimated from experimental measurements with these 11 sequences, by cloning promoter sequences in front of YFP to measure expression level and in front of *URA3* to measure fitness. Unless otherwise noted, yeast were grown at 30 °C, in an orbital shaker incubator at 225 rpm. Each vector was transformed into yeast (S288C::*ura3*), and three independent transformants were selected per vector to serve as biological replicates. For measuring expression, yeast were grown overnight in either YPD+NAT (yeast extract, peptone, dextrose, with 75 μg ml^{-1} nourseothricin) or SD-Ura (Sunrise Science 1703-500), and then re-inoculated in the morning and allowed to grow for 6 h before measuring expression by flow cytometry for each replicate as the log ratio of YFP to the constant background red fluorescent protein (RFP), including only cells obtaining the top 50% of RFP expression. Fitness was obtained by measuring the growth rate of each yeast strain in either SD-Ura or YPD+NAT+5-FOA (0.25 mg ml^{-1} 5-FOA). Yeast were grown continuously in triplicate in log phase, with linear shaking at 30 °C in a Synergy H1 plate reader (Biotek), by diluting each well to maintain OD < 0.7, with OD measured at 15-min intervals. Growth rate was defined for each replicate as the median of the instantaneous smoothed growth rates over five measurements in log phase, considering only time points at which 0.05 < OD < 0.5. The expression and growth rate of each promoter were summarized as the mean of the three replicates.

Characterizing trajectories under conflicting expression objectives in different environments

Simulations of sequence evolution in two complementary environments with opposite selective pressures (defined medium and complex medium) were initialized with a set of native yeast promoter sequences (present at −160 to −80 relative to the TSS) as the starting generation 0, with the objective function defined as the difference in predicted expression between defined and complex medium (Fig. 2h, Extended Data Fig. 2d–g) using convolutional models trained in the respective medium. The difference in expression between the two conditions was maximized at each iteration, which assumes that the cells are exposed to both environments before the mutations can reach fixation, an example of evolution in rapidly fluctuating environments[31]. For simplicity, it is assumed that fitness is directly proportional to higher expression in one condition and to lower expression in the other, such that mutations will be considered favourable even if they decrease fitness in one condition so long as they increase it in the other condition by a greater amount.

One simulation aimed to maximize the expression difference (defined minus complex), and the other to minimize it (maximizing complex minus defined). For each sequence in generation n, the sequence from its 3L mutational neighbourhood that had the maximum (or separately, minimum) value for the objective function based on the convolutional model prediction is picked for generation $n+1$, to a total of 10 generations. The simulations were subsequently repeated using the transformer model, yielding similar results (Supplementary Fig. 17b–f).

Motifs that were enriched in the sequences of generation 10 compared to the starting sequences were identified de novo using DREME[87], and each of the top five consensus motifs were used as queries to search the YeTFaSCo database[88], reporting the closest match, or one of multiple similar matches.

Finding orthologous promoters in the 1,011 *S. cerevisiae* genomes dataset

To identify orthologues of S288C promoters in the whole-genome sequences of the 1,011 yeast strains[37], BLAT[89] was used to identify regions of ≥80% identity with each −160 to −80 region (relative to the TSS) annotated in the reference S288C genome sequence (R64)[90]. Any strains with more than one such match, where the match contained insertions or deletions, or had incomplete matches, were excluded on a gene-by-gene basis. Genes with more than 1.2 matches with ≥80% identity per genome, on average, were excluded altogether.

Computing the ECC

To calculate the ECC (a regulatory analogue[2,35,91,92] of d_N/d_S[34,93,94]), for each yeast gene promoter, the transformer model was used to predict an expression value for each orthologous promoter in the 1,011 yeast genomes (above), defining an expression distribution with a standard deviation σ_B. Next, a set of sequences with random mutations was generated from each gene's consensus promoter sequence (defined as the most abundant base at each 'position across the strains), such that the number of sequences at each Hamming distance from the consensus promoter sequence was the same for the natural and simulated sets. Here, mutations introduced to create random variation sampled each base with equal probability; using observed mutation rates yielded similar results (Supplementary Information). The same transformer model was used to predict the expression of the simulated sequences, and to calculate its standard deviation σ_C. The nominal ECC is $\log(\sigma_C/\sigma_B)$. Because the variance on simulated sequences is better estimated than in natural orthologues (the sequences of which may be more constrained), a constant correction factor is subtracted, calculated by creating a

second simulated set of randomly mutated sequences the diversity of which is limited to the same extent as in the natural set, by creating only one random mutation for every unique sequence in the set of native orthologues. Finally, the expression for this second set of sequences is predicted by the transformer model, and its standard deviation ($\sigma_{C'}$) is used to calculate a null ECC for each gene ($\log(\sigma_C/\sigma_{C'})$); the median of these null ECCs over all the genes is used as the constant correction factor.

$$C = \text{median}_{\forall \text{ genes},i}\left(\log_2\left(\frac{\sigma_{C_i}}{\sigma_{C'_i}}\right)\right)$$

(An extensive description of the correction factor is provided in the 'ECC calculation details and considerations' section of the Supplementary Information.)

The corrected ECC for gene g is then:

$$\text{ECC}_g = \log_2\left(\frac{\sigma_{C_g}}{\sigma_{B_g}}\right) - C.$$

The computed ECC values for all yeast genes, available in Supplementary Table 1, were used to identify cases of presumed stabilizing selection (selection favouring a fixed non-extreme value of a trait), diversifying (disruptive) selection (selection favouring more than one extreme values of a trait; as opposed to a single fixed intermediate value) and directional (positive) selection (selection favouring a single extreme value of a trait over all other possible values of the trait). Recomputing the ECC values for all yeast genes using the S288C reference sequences instead of the consensus sequence for the promoters of each gene yielded very similar results.

In addition to each ECC value, a z-score and P values for the confidence that the observed ECC values differ from neutrality were also calculated. For each gene's true ECC, a set of matched random ECC values were calculated, in which the denominator is a set of sequences matched for Hamming distance distribution and the total number of unique sequences. The null ECC mean and standard deviation were calculated from 1,111 such simulations, and used to calculate a z-score for how extreme the actual ECC would be under this null distribution. This z-score acts as a signed P value (negative representing divergent expression and positive representing conservation) from which P values are obtained (using the 'scipy.stats.norm.sf' function on the absolute value of the z-score in Scipy[95] and multiplying the function's output by 2 to get a two-sided P value) (Supplementary Table 1).

Inferring expression conservation across *Saccharomyces* species using RNA-seq data and comparing with ECC values

Published reads per kilobase million (RPKM) values for orthologues of *S. cerevisiae* genes in closely related *Saccharomyces* species[41] were obtained from the GEO (accession GSE83120). Only genes for which expression was quantified in all species were used in subsequent analysis. RPKM values were $\log_2$-scaled after adding a pseudo count of 2, and the variance in expression of each gene across the species was calculated. Genes were ranked by their gene expression variance, and the 2% of genes with the lowest variance were considered as having conserved gene expression levels ('expression conserved'), whereas the 2% with the highest variance were considered 'expression not-conserved'. The analysis was robust to the choice of thresholds (Supplementary Information). The P value of a two-sided Wilcoxon rank-sum test was computed by comparing the ECC values for genes in the 'expression conserved' and 'expression not-conserved' categories (implemented using the 'scipy.stats.ranksums' SciPy[95] function). To control for the dependence between expression mean and variance, the analysis was repeated using the coefficient of variation ($P = 1.05 \times 10^{-4}$) and the coefficient of dispersion ($P = 2.42 \times 10^{-4}$) instead of variance, yielding similar results.

Experimental protocol for RNA-seq measurements from 11 Ascomycota species

RNA-seq was performed on samples from the following 11 Ascomycota yeast species: *Saccharomyces cerevisiae*, *Saccharomyces bayanus*, *Naumovozyma* (*Saccharomyces*) *castellii*, *Candida glabrata*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, *Candida albicans*, *Yarrowia lipolytica*, *Schizosaccharomyces japonicus*, *Schizosaccharomyces octosporus* and *Schizosaccharomyces pombe*. Each of the 11 species was grown in BMW medium, chosen to minimize cross-species growth differences, as previously described[96]. *N. castellii* was grown at 25 °C whereas the other species were grown at 30 °C. RNeasy Midi or Mini Kits (Qiagen) were used to isolate total RNA from log-phase cells by mechanical lysis using the manufacturer's instructions as previously described[96]. dUTP strand-specific RNA-seq libraries were constructed as previously described[97] with the following modifications. (1) The polyA$^+$-selected RNA was fragmented in a 40-µl reaction containing 1× fragmentation buffer (Affymetrix) by heating at 80 °C for 4 min followed by clean-up by ethanol precipitation for all libraries (except *Y. lipolytica*, *S. pombe*, *S. japonicus* and *S. octosporus*; for these species, the conditions described previously were used[97]), followed by clean-up using 1.8× RNAClean XP beads (Beckman Coulter Genomics). (2) For *C. glabrata*, *K. lactis*, *S. bayanus*, *S. pombe*, *S. japonicus* and *S. octosporus* libraries, the adapter ligation was performed overnight at 16 °C. For the rest, this was done at 16 °C for 2 h as described previously[97]. (3) Normalization was carried out based on the cDNA input and pooling of selected Illumina barcoded-adaptor-ligated cDNA products followed by gel size selection as follows: range of 275 to 575 bp for pooled *C. albicans*, *K. waltii* and *N. castellii* libraries, and 375 to 575 bp for *C. glabrata*, *K. lactis* and *S. bayanus* libraries. For the other libraries, no pooling was performed before gel size selection—range of 310 to 510 bp for *Y. lipolytica* and 350 to 550 bp for *S. pombe*, *S. japonicus* and *S. octosporus*. (4) The final PCR product was purified by 1.8× AMPure XP beads (Beckman Coulter Genomics) followed by a second gel size selection for the range of 300 to 575 bp for *C. albicans*, *K. waltii* and *S. castellii* libraries, but no second gel size selection was performed for the other libraries. The pooled final library was sequenced on 1 to 4 lanes of a HiSeq2000 (Illumina) with 68 base (*Y. lipolytica* had 76 base) paired-end reads and 8 base index reads.

Transcript assembly, mapping and expression calculation for RNA-seq in 11 Ascomycota species

For each of the 11 Ascomycota yeast species above, reads were assembled using Trinity[98](version 'trinityrnaseq_r2012-05-18') and the assembled transcripts were mapped onto the assemblies to the respective genomes using GMAP[99]. The Jaccard coefficient was used to join adjacent assemblies given enough connecting reads (using the Trinity default of 0.35 for the Jaccard cut-off). Finally, after mapping all assembled transcripts, the Jaccard coefficient was used to clip assemblies that did not have enough support over a certain region. For each of the species, assembled transcripts were mapped to the genome sequence[100] using BLAT[89]. Estimated expression values were calculated for each transcript using RSEM[101] (defined in RSEM as the estimate of the number of fragments that are derived from a given isoform or gene, or the expectation of the number of alignable and unfiltered fragments that are derived from an isoform or gene given the maximum likelihood abundances). Only reads mapping to the sense mRNA strand were considered. Orthology between genes in different species was used as previously described[100].

Inferring expression conservation across Ascomycota species using RNA-seq data and comparing with ECC values

Estimated expression values from the 11 Ascomycota species RNA-seq data were used after removing all genes with missing values in expression for more than three species. Estimated expression values were $\log_2$-scaled after adding a pseudo count of 1, and the variance in

expression for each gene across the species was calculated. Genes were ordered by their variance in expression across the reported fungal species. Here, the 10% of genes with the lowest expression variance were considered to have 'conserved' expression, and the 10% with highest expression variance were considered to have expression 'not conserved'. The analysis was robust to the choice of thresholds (Supplementary Information). The *P* value of a two-sided Wilcoxon rank-sum test was computed by comparing the ECC values for genes in the 'conserved' and 'not conserved' categories (implemented using the 'scipy.stats.ranksums' SciPy[95] function). Similar results were obtained when repeating the analysis using the coefficient of variation ($P = 4.22 \times 10^{-5}$) and the coefficient of dispersion ($P = 8.05 \times 10^{-5}$) instead of variance.

Inferring expression conservation across mammalian species using RNA-seq data and comparing with ECC values

Ensembl Biomart[102] was used to find one to one orthologues of *S. cerevisiae* genes in humans (of 'Human homology type' 'ortholog_one2one'; all 'ortholog_one2many' and 'many2many' orthologues were excluded). For these human orthologues of yeast genes, the previously reported 'evolutionary variance' values across mammalian species from the original publication[42] (based on an Ornstein Uhlenbeck (OU model)[42]) were directly used. Here, the 25% of genes with the lowest 'evolutionary variance' were considered to have conserved expression and the top 25% were considered to be not conserved (the same thresholds used in the original study[42]). The analysis was robust to the choice of thresholds (Supplementary Information). This was done separately for each profiled tissue (brain, heart, kidney, liver, lung and skeletal muscle). Subsequently, a human orthologue for a yeast gene was considered to have conserved (or non-conserved) expression if it was found to have conserved (or non-conserved) expression in at least one of the profiled tissues. Genes with conflicting expression conservation classes across tissues were excluded from the analysis. The *P* value of a two-sided Wilcoxon rank-sum test was computed by comparing the ECC values for genes in the 'conserved' and 'not conserved' categories (implemented using the 'scipy.stats.ranksums' SciPy[95] function).

Quantifying sequence dissimilarity using mean Hamming distance

For each group of orthologous yeast gene promoters (with ungapped alignments), the mean of Hamming distances between each pair of orthologous promoters across the 1,011 isolates was calculated.

Generation of *CDC36* promoter strains by allele swapping

Strains with a restored Upc2p-binding site in the *CDC36* promoter region were obtained using a previously described CRISPR–Cas9 method[103]. Guide RNAs (gRNAs) were designed using the Benchling online tool (https://www.benchling.com/) and cloned in a pGZ110 derived plasmid[104], using standard 'Golden Gate Assembly'[105]. Plasmids carrying the gRNA and Cas9 gene were then co-transformed with a synthetic DNA fragment (ssODN) composed of a 100-bp sequence with perfect complementarity to the background promoter sequence (WE) but for the centrally located targeted alleles that overlap the Upc2p-binding site. Allele swapping was confirmed by Sanger sequencing (Macrogen). Sequences were analysed using the SGRP (Saccharomyces Genome Resequencing Project) BLAST server (http://www.moseslab.csb.utoronto.ca/sgrp/blast_new/) and the MUSCLE tool in Geneious v.10.1. All primers and ssODNs used are listed in Supplementary Table 2.

RNA extraction and qPCR of *CDC36*

Gene expression analysis was performed by qPCR from cultures grown in SD medium supplemented with uracil (0.02% p/v). Samples were grown until exponential phase (OD 0.6–0.8), collected by centrifugation and treated with 10 units of Zymolyase 20T (50 mg ml^{-1}) for 30 min at 37 °C. RNA was extracted using E.Z.N.A Total RNA kit I (OMEGA) according to the manufacturer's instructions. Genomic DNA traces were then removed by treating samples with DNase I (Promega). RNA concentrations were estimated using a Qubit system and verified by 1.5% agarose gel. RNA extractions were performed in three biological replicates.

cDNA was synthesized using 200 units of M-MLV Reverse transcriptase (Promega), 0.5 µg of Oligo (dT)15 primer and 1 µg of RNA in a final volume of 25 µl according to the manufacturer's instructions. qPCR reactions were carried out using Brilliant II SYBR Green QPCR Master Mix (Agilent Technologies) in a final volume of 10 µl, containing 0.2 µM of each primer and 1 µl of the cDNA previously synthesized. qPCR reactions were carried out in three technical replicates per biological replicate using an Eco Real-Time PCR system (Illumina) under the following conditions: 95 °C for 15 min and 40 cycles at 95 °C for 10 s and 58 °C for 30 s. Primers used are listed in Supplementary Table 2. The relative expression of *CDC36* was quantified using the $2(-\Delta\Delta C_t)$ approach[106], and normalized with two housekeeping genes as previously described[107], using the median C_t of the three technical replicates for each sample. The housekeeping genes *ACT1* and *RPN2* were used as previously described[108].

Growth curves of *CDC36* mutant and wild-type alleles

Growth curves incorporating carbon source switching from glucose to galactose were generated as previously described[109]. Pre-cultures were grown in YNB containing 5% glucose medium at 30 °C for 24 h. Cultures were then diluted to an initial optical density at 600 nm ($OD_{600\,nm}$) of 0.1 in fresh YNB 5% glucose medium for an extra overnight growth. The next day, cultures were used to inoculate a 96-well plate with a final volume of 200 µl YNB with 5% galactose with an initial $OD_{600\,nm}$ of 0.1. In parallel, a control plate containing YNB with 5% glucose was similarly inoculated. All experiments were performed in triplicate. $OD_{600\,nm}$ was monitored every 30 min using a Tecan Sunrise absorbance microplate reader (Tecan Group). The kinetic parameters of lag phase, growth efficiency ($\Delta OD_{600\,nm}$) and maximum specific growth rate (μ_{max}) were determined as previously described[110], fitting the curves with the Gompertz function using R v.3.3.2. All growth parameters are expressed as the ratio of growth within YNB + galactose to YNB + glucose to control for phenotypic variation that results from something other than the carbon source switch.

Fitness responsivity

The empirically determined relationships between the expression levels and organismal fitness for each of 80 genes[11] were re-analysed. Published expression-to-fitness curves in glucose medium for each of 80 genes were obtained from the Supplementary Data of the original publication[11]. For each of these curves, the total variation (Extended Data Fig. 5) was calculated by partitioning the expression range into 36 regular intervals (as reported in the 'impulse fit' of the expression-to-fitness curves in the original publication[11]) and summing the absolute difference in fitness at the endpoints of each partition as follows $\sum |F_{GENE}(e_{i+1}) - F_{GENE}(e_i)|$ for each gene's expression-to-fitness function, $F_{GENE}(e)$. The same qualitative relationship between a gene's ECC and fitness responsivity as reported in other studies[111–113] was observed, including *LCB2* (ECC 2.15 and high fitness responsivity[112]) and *MLS1* (ECC −1.32 and extremely low fitness responsivity[113]).

Mutational robustness

For every sequence, mutational robustness was defined as the fraction of sequences in its 3*L* mutational neighbourhood that altered the expression by an amount less than ϵ, where ϵ is set at two times the standard deviation of expression variance across all genes with an ECC > 0 (here, $\epsilon = 0.1616$; ECC calculated using the 1,011 *S. cerevisiae* genomes; Extended Data Fig. 4c). Using different values for ϵ yielded very similar results.

The evolvability vector

To derive the evolvability vector for a given sequence, expression changes associated with single base changes in every possible position

were sorted to obtain a monotonically increasing vector of length $3L$ for each sequence of length L (here, $L = 80$; $3L = 240$; Fig. 4a, Methods). Formally, to compute an evolvability vector for a sequence s_0, for each sequence s_i in the $3L$ mutational neighbourhood of s_0, the difference between the predicted expression of s_i and that of s_0: $d_i = f(s_i) - f(s_0)$ was calculated, in which $f(s)$ represents the predicted expression of the transformer model. The evolvability vector is defined as the vector $\mathbf{D}$ ($\{d_1, d_2, \ldots, d_{3L}\}$), sorted such that $d_i \geq d_{i-1}, \forall\ i$ (that is, d_i values are in ascending order).

Power law distribution analysis

The list of the absolute values from the evolvability vectors for all native sequences was used to define the distribution of the magnitude of the expression effect of mutations. The powerlaw[114] Python package was used to determine whether the data fit a power law distribution. The 'Fit' function with an 'xmin' parameter of 0.5 was used to determine the exponent and the 'distribution_compare' function was used to determine the P value for the fit (Extended Data Fig. 4d, Supplementary Fig. 2h).

Characterizing the archetypal evolvability space

The evolvability vectors for a new random sample of a million sequences were used as input to an autoencoder with an archetypal regularization constraint[45] on the embedding layer. The autoencoder was trained using the AANet implementation made available with the publication[45] with no noise added to the archetypal layer during training, a linear activation on the output layer, an equal weight of 1 on each of the loss terms (the mean squared error loss term along with the non-negativity and convexity constraints), a learning rate of 0.001 and a minibatch size of 4,096. The autoencoder accepts an evolvability vector (of length 240 for an 80-bp sequence) as input to the first encoder layer, in which each node in the input layer is connected to each node in the encoder layer (fully connected layer). Every layer in the autoencoder was fully connected. The encoder architecture used was[1024, 512, 256, 128, 64], in which each entry corresponds to the number of nodes in the corresponding hidden layer and the decoder architecture was the encoder's mirror image. The output layer was the same shape as input layer and each node in the last decoder layer was connected to each node in the output layer. To select the optimal number of archetypes, the autoencoder was first trained for a 1,000 minibatches separately for 1 to 9 archetypes. Following the recommended approach[45] for picking the optimal number of archetypes, we used an elbow plot of mean squared error on the evolvability vectors (here, using native sequences) versus the number of archetypes in the autoencoder (Extended Data Fig. 6a).

The autoencoder was then trained from scratch with 3 archetypes, using the full training data and parameters for 250,000 batches. As this autoencoder aims to reconstruct the original evolvability vector for each sequence by learning feature representations after passing them through an information bottleneck, its reconstruction accuracy was first verified on the set of native yeast promoter sequences (Extended Data Fig. 6b; Pearson's $r = 0.992$). To visualize the evolvability vectors for sequences in 2 dimensions (2D), the evolvability vectors corresponding to the three archetypes were first generated by decoding their archetypal latent space coordinates ((1,0,0), (0,1,0) and (0,0,1)) through the decoder, and MDS was performed on the decoded evolvability vectors of the archetypes. Then, as previously described[45], the encoded evolvability vector of each new sequence was projected into the 2D MDS space by representing it as a mixture of the archetypes and interpolating them between the MDS coordinates of each archetype. For every sequence, the following equivalent representations can now be computed: (i) its evolvability vector; (ii) an archetypal triplet quantifying the similarity of its encoded (latent space) evolvability vector to the three archetypes; and (iii) a two-dimensional MDS coordinate[45] for visualizing the evolvability vectors. The representation of the evolvability vector for each sequence in this archetypal space is now bounded by a simplex (the vertices of which correspond to the three evolvability archetypes). For each native and natural yeast promoter sequence from the sequence space, the archetypal triplet and MDS coordinates were inferred using its evolvability vector with this trained autoencoder. The MDS coordinates for the archetypes and the native yeast promoter sequences were used to generate the visualizations of the sequence space shown. This archetypal characterization of evolvability vectors allows the encoding and visualization of sequences by their evolvability in the context of a fitness landscape.

Visualizing promoter fitness landscapes

A total of 1,000 random sequences were sampled and projected onto the MDS coordinate system for visualizing the sequence space described above. The expression level of each sequence was calculated using our model, and expression values were scaled so that the minimum was 0 and maximum was 1. Previously quantified expression-to-fitness relationships[11] to compute fitness (fraction of wild-type growth rate) by using cubic spline interpolation (implemented using the 'scipy.interpolate.CubicSpline' SciPy[95] function) on the expression level after scaling the measured expression-to-fitness curves to have an expression range of 0 to 1. These fitness values were then used to generate the contour plots (implemented using the 'matplotlib.pyplot.tricontourf' function; Fig. 4e, Extended Data Fig. 7) that visualize the fitness landscape in that gene's promoter sequence space.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data generated for this study are available at the NCBI GEO with accession numbers GSE163045 and GSE163866. All models and processed data are available on Zenodo at https://zenodo.org/record/4436477.

Code availability

Code is available on GitHub at https://github.com/1edv/evolution and CodeOcean at https://codeocean.com/capsule/8020974/tree. A web app is available at https://1edv.github.io/evolution/.

51. Kosuri, S. et al. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl Acad. Sci. USA.* **110**, 14024–14029 (2013).
52. Shalem, O. et al. Systematic dissection of the sequence determinants of gene 3′ end mediated expression control. *PLoS Genet.* **11**, e1005147 (2015).
53. Kinney, J. B., Murugan, A., Callan, C. G. Jr & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci. USA.* **107**, 9158–9163 (2010).
54. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
55. Melnikov, A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
56. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc. Natl Acad. Sci. USA* **109**, 19498–19503 (2012).
57. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
58. Townsley, K. G., Brennand, K. J. & Huckins, L. M. Massively parallel techniques for cataloguing the regulome of the human brain. *Nat. Neurosci.* **23**, 1509–1521 (2020).
59. Renganaath, K. et al. Systematic identification of cis-regulatory variants that cause gene expression differences in a yeast cross. *eLife* **9**, e62669 (2020).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
61. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
62. Travers, C. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
63. Avsec, Ž. et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
64. Quang, D. & Xie, X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* **166**, 40–47 (2019).

65. Zhou H. et al. Towards a better understanding of reverse-complement equivariance for deep learning models in genomics. *Proc. 16th Machine Learning in Computational Biology meeting* **165**, 1–33 (2022).
66. Morrow, A. et al. Convolutional kitchen sinks for transcription factor binding site prediction. Preprint at https://arxiv.org/abs/1706.00125 (2017).
67. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
68. Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P. & Paul, S. B. Global importance analysis: an interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Comput. Biol.* **17**, e1008925 (2021).
69. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).
70. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *International Conference on Learning Representations* (Poster) (2015).
71. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogenous systems. Software available from https://www.tensorflow.org/ (2015).
72. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proc. 44th Annual International Symposium on Computer Architecture* 1–12 (2017).
73. Li, J., Pu, Y., Tang, J., Zou, Q. & Guo, F. DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. *Brief. Bioinform.* **22**, bbaa159 (2020).
74. Ullah, F. & Ben-Hur, A. A self-attention model for inferring cooperativity between regulatory features. *Nucleic Acids Res.* **49**, e77 (2021).
75. Clauwaert, J., Menschaert, G. & Waegeman, W. Explainability in transformer models for functional genomics. *Brief. Bioinform.* **22**, bbab060 (2021).
76. Hinton, G. & Tieleman, T. Lecture 6.5—RmsProp: divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning* **4**, 26–31 (2012).
77. Sinai, S. et al. AdaLead: a simple and robust adaptive greedy search algorithm for sequence design. Preprint at https://arxiv.org/abs/2010.02141 (2020).
78. Linder, J., Bogard, N., Rosenberg, A. B. & Seelig, G. A generative neural network for maximizing fitness and diversity of synthetic DNA and protein sequences. *Cell Syst.* **11**, 49–62 (2020).
79. Brookes, D., Park, H. & Listgarten, J. Conditioning by adaptive sampling for robust design. *Proc. Mach. Learn. Res.* **97**, 773–782 (2019).
80. Killoran, N., Lee, L. J., Delong, A., Duvenaud, D. & Frey, B. J. Generating and designing DNA with deep generative models. *Neurips Computational Biology Workshop* (2017).
81. Fortin, F.-A., Rainville, F.-M. D., Gardner, M.-A., Parizeau, M. & Gagné, C. DEAP: evolutionary algorithms made easy. *J. Mach. Learn. Res.* **13**, 2171–2175 (2012).
82. Jaeger, S. A. et al. Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites. *Genomics* **95**, 185–195 (2010).
83. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
84. Sniegowski, P. D. & Gerrish, P. J. Beneficial mutations and the dynamics of adaptation in asexual populations. *Phil. Trans. R. Soc. B* **365**, 1255–1263 (2010).
85. Szendro, I. G., Franke, J., de Visser, J. A. & Krug, J. Predictability of evolution depends nonmonotonically on population size. *Proc. Natl Acad. Sci. USA* **110**, 571–576 (2013).
86. Orr, H. A. The population genetics of adaptation: the adaptation of DNA Sequences. *Evolution* **56**, 1317–1330 (2002).
87. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
88. de Boer, C. G. & Hughes, T. R. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.* **40**, D169–D179 (2012).
89. Kent, W. J. BLAT—the BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
90. Cherry, J. M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
91. Smith, J. D., McManus, K. F. & Fraser, H. B. A novel test for selection on *cis*-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. *Mol. Biol. Evol.* **30**, 2509–2518 (2013).
92. Liu, J. & Robinson-Rechavi, M. Robust inference of positive selection on regulatory sequences in the human brain. *Sci. Adv.* **6**, eabc9863 (2020).
93. Rice, D. P. & Townsend, J. P. A test for selection employing quantitative trait locus and mutation accumulation data. *Genetics* **190**, 1533–1545 (2012).
94. Denver, D. R., Morris, K., Lynch, M. & Thomas, W. K. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430**, 679–682 (2004).
95. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
96. Thompson, D. A. et al. Evolutionary principles of modular gene regulation in yeasts. *eLife* **2**, e00603 (2013).
97. Yassour, M. et al. Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol.* **11**, R87 (2010).
98. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
99. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
100. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61 (2007).
101. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
102. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
103. DiCarlo, J. E. et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR–Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013).
104. Fleiss, A. et al. Reshuffling yeast chromosomes with CRISPR/Cas9. *PLoS Genet.* **15**, e1008332 (2019).
105. Horwitz, A. A. et al. Efficient multiplexed integration of synergistic alleles and metabolic pathways in yeasts via CRISPR–Cas. *Cell Syst.* **1**, 88–96 (2015).
106. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).
107. Vandesompele, J. et al. Accurate normalization of real-time quantitative RT–PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research0034.1 (2002).
108. Teste, M.-A., Duquenne, M., François, J. M. & Parrou, J.-L. Validation of reference genes for quantitative expression analysis by real-time RT–PCR in *Saccharomyces cerevisiae*. *BMC Mol. Biol.* **10**, 99 (2009).
109. Mardones, W. et al. Rapid selection response to ethanol in *Saccharomyces eubayanus* emulates the domestication process under brewing conditions. *Microb. Biotechnol.* https://doi.org/10.1111/1751-7915.13803 (2021).
110. Ibstedt, S. et al. Concerted evolution of life stage performances signals recent selection on yeast nitrogen use. *Mol. Biol. Evol.* **32**, 153–161 (2015).
111. Rich, M. S. et al. Comprehensive analysis of the SUL1 promoter of *Saccharomyces cerevisiae*. *Genetics* **203**, 191–202 (2016).
112. Rest, J. S. et al. Nonlinear fitness consequences of variation in expression level of a eukaryotic gene. *Mol. Biol. Evol.* **30**, 448–456 (2013).
113. Bergen, A. C., Olsen, G. M. & Fay, J. C. Divergent MLS1 promoters lie on a fitness plateau for gene expression. *Mol. Biol. Evol.* **33**, 1270–1279 (2016).
114. Alstott, J., Bullmore, E. & Plenz, D. Powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS One* **9**, e85777 (2014).

Acknowledgements We thank Google TPU Research Cloud for TPU access, L. Gaffney for help with figure preparation, Broad Genomics Platform for sequencing work, J.-C. Hütter for advice on fitness responsivity, J. Pfiffner-Borges for help with RNA-seq, R. Yu, B. Lee and N. Jaberi for manuscript feedback and members of the A.R. laboratory for discussions. E.D.V. was supported by the MIT Presidential Fellowship; C.G.d.B. was supported by a Canadian Institutes for Health Research Fellowship and the NIH (K99-HG009920-01); and F.A.C. and J.M. were supported by ANID (Programa Iniciativa Científica Milenio, ICN17_022). Work was supported by the Klarman Cell Observatory, Howard Hughes Medical Institute (HHMI) and Google TPU Research Cloud (https://sites.research.google/trc/about/). A.R. was an Investigator of the HHMI.

Author contributions E.D.V., C.G.d.B. and A.R. conceived, designed and supervised the study. E.D.V. and C.G.d.B. performed the analyses. M.Y., L.F., X.A. and D.A.T. performed and D.A.T., J.Z.L. and A.R. supervised the Ascomycota cross-species RNA-seq experiments. J.M. performed and F.A.C. supervised the *CDC36* experiments. E.D.V. and C.G.d.B. performed the rest of the experiments. E.D.V., C.G.d.B. and A.R. wrote the manuscript.

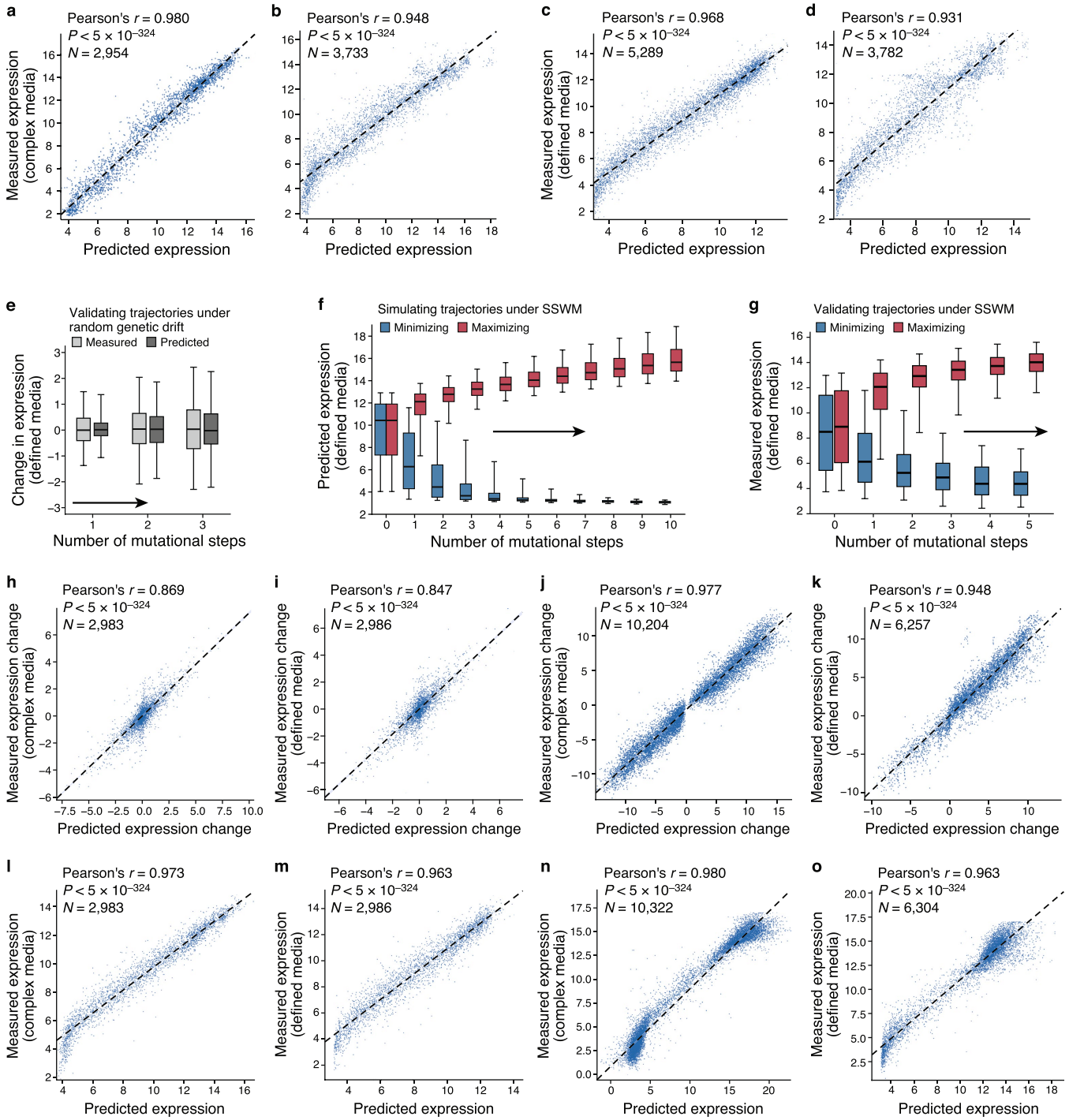
Competing interests A.R. is a co-founder and equity holder of Celsius Therapeutics and Immunitas and until 31 July 2020 was a member of the scientific advisory board of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov. As of 1 August 2020, A.R. is an employee of Genentech. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04506-6.

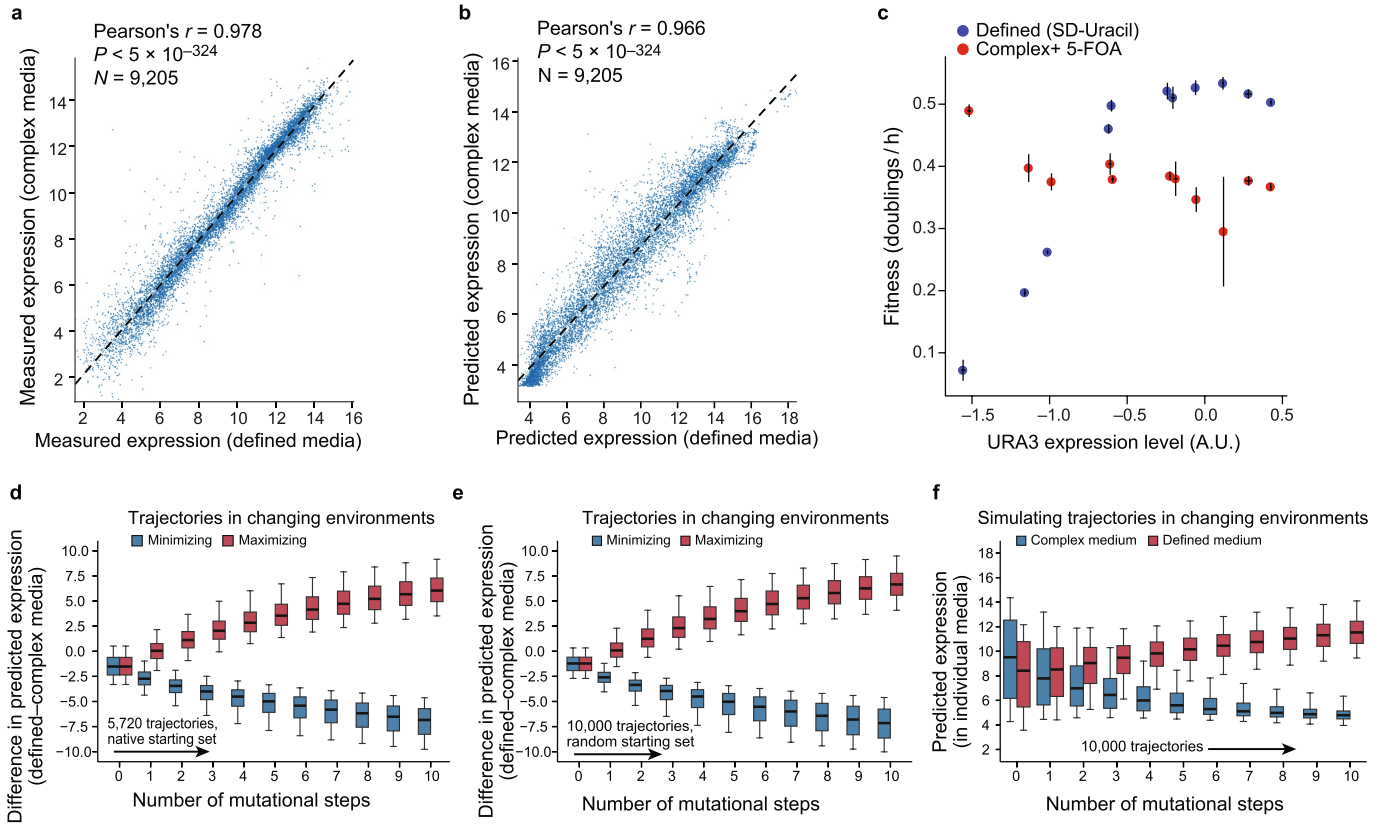
Correspondence and requests for materials should be addressed to Eeshit Dhaval Vaishnav, Carl G. de Boer or Aviv Regev.

Peer review information *Nature* thanks Martin Taylor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



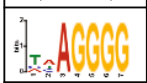
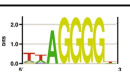
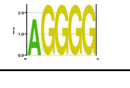
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | The convolutional sequence-to-expression model generalizes reliably and characterizes sequence trajectories under different evolutionary regimes. a–**d**, Prediction of expression from sequence in complex (YPD) (**a**, **b**) and defined (SD-Uracil) (**c**, **d**) medium. Predicted (*x* axis) and experimentally measured (*y* axis) expression for (**a**, **c**) random test sequences (sampled separately from and not overlapping with the training data) and (**b**, **d**) native yeast promoter sequences containing random single base mutations. Top left: Pearson's *r* and associated two-tailed *P* value. Compression of predictions in the lower left results from binning differences during cell sorting in different experiments (Supplementary Notes). **e**, Experimental validation of trajectories from simulations of random genetic drift. Distribution of measured (light grey) and predicted (dark grey) changes in expression in the defined medium (SD-Uracil) (*y* axis) for the synthesized randomly designed sequences (n = 2,986) at each mutational step (*x* axis). Midline: median; boxes: interquartile range; whiskers: 5th and 95th percentile range. **f**, **g**, Simulation and validation of expression trajectories under SSWM in defined medium (SD-Uracil). **f**, Distribution of predicted expression levels (*y* axis) in defined medium at each evolutionary time step (*x* axis) for sequences under SSWM favouring high (red) or low (blue) expression, starting with native promoter sequences (n = 5,720). Midline: median; boxes: interquartile range; whiskers: 5th and 95th percentile range. **g**, Experimentally measured expression distribution in defined medium (*y* axis) for the synthesized sequences (n = 6,304 sequences; 637 trajectories) at each mutational step (*x* axis) from predicted mutational trajectories under SSWM, favouring high (red) or low (blue) expression. Midline: median; boxes: interquartile range; whiskers: 5th and 95th percentile range. **h**–**o**, Experimental validation of predicted expression for sequences from the random genetic drift and SSWM simulations. Experimentally measured (*y* axis) and predicted (*x* axis) expression level (**l**–**o**) or expression change from the starting sequence (**h**–**k**) in complex (**h**, **j**, **l**, **n**) or defined (**i**, **k**, **m**, **o**) medium using sequences from the random genetic drift (Fig. 2e, Extended Data Fig 1e, h, i, l, m here) and SSWM (Fig. 2g, Extended Data Fig 1g, j, k, n, o here) validation experiments. Top left: Pearson's *r* and associated two-tailed *P* values.

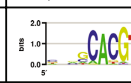
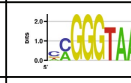
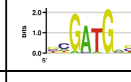
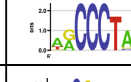
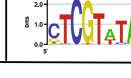


g

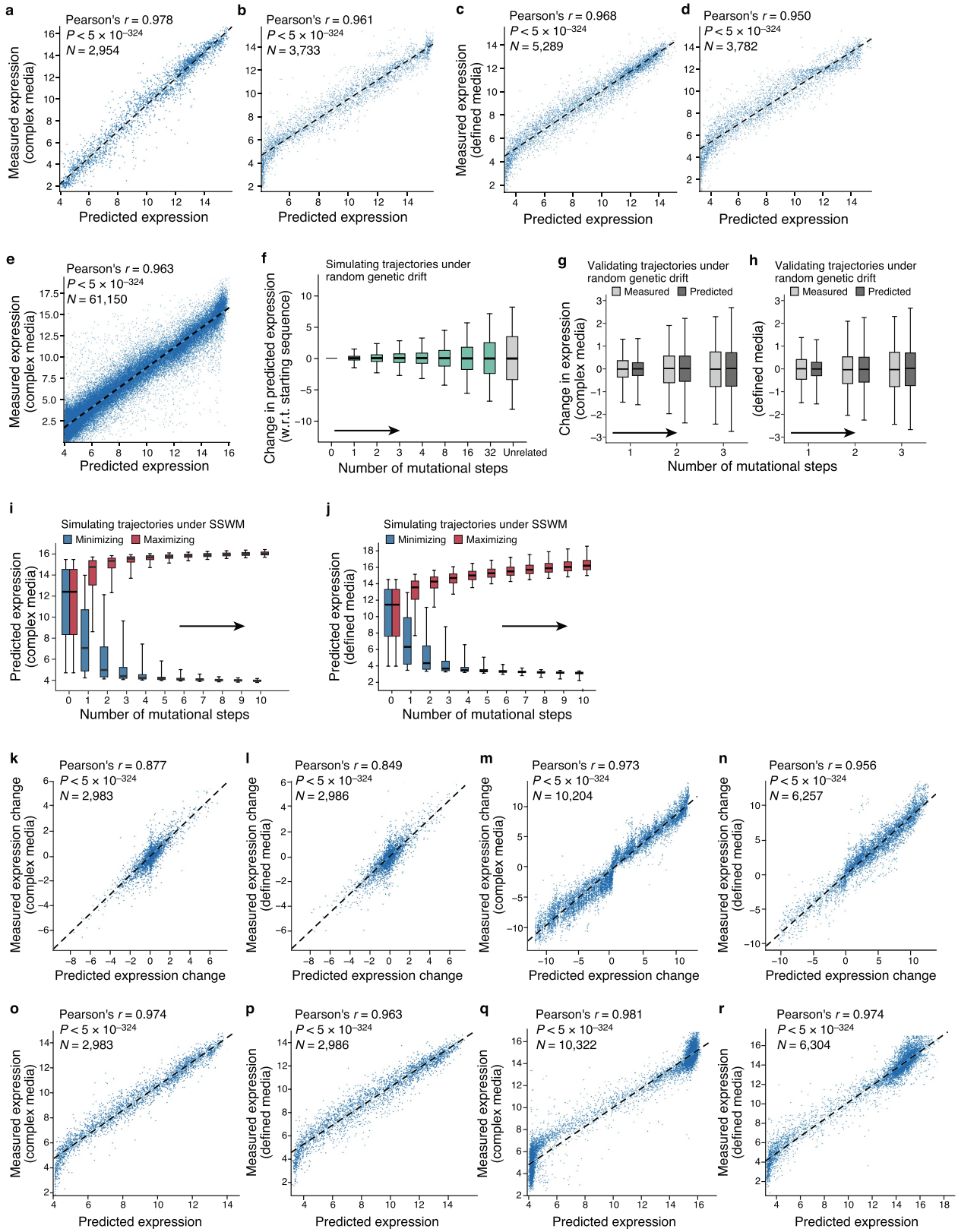
Top 5 motifs enriched in sequences evolved to **MAXIMIZE** difference in expression between the defined and complex media

Found motif (DREME)	DREME *E*-value	Rank in native	Known factor	Known motif (YeTFaSCo)
	3.1×10^{-1953}	1	Several (e.g., RPH1)	
	1.9×10^{-269}	2	INO2/4	
	1.1×10^{-218}	3	None	
	9.8×10^{-124}	24	None	
	1.7×10^{-91}	8	Several (e.g. GIS1)	

Top 5 motifs enriched in sequences evolved to **MINIMIZE** difference in expression between the defined and complex media

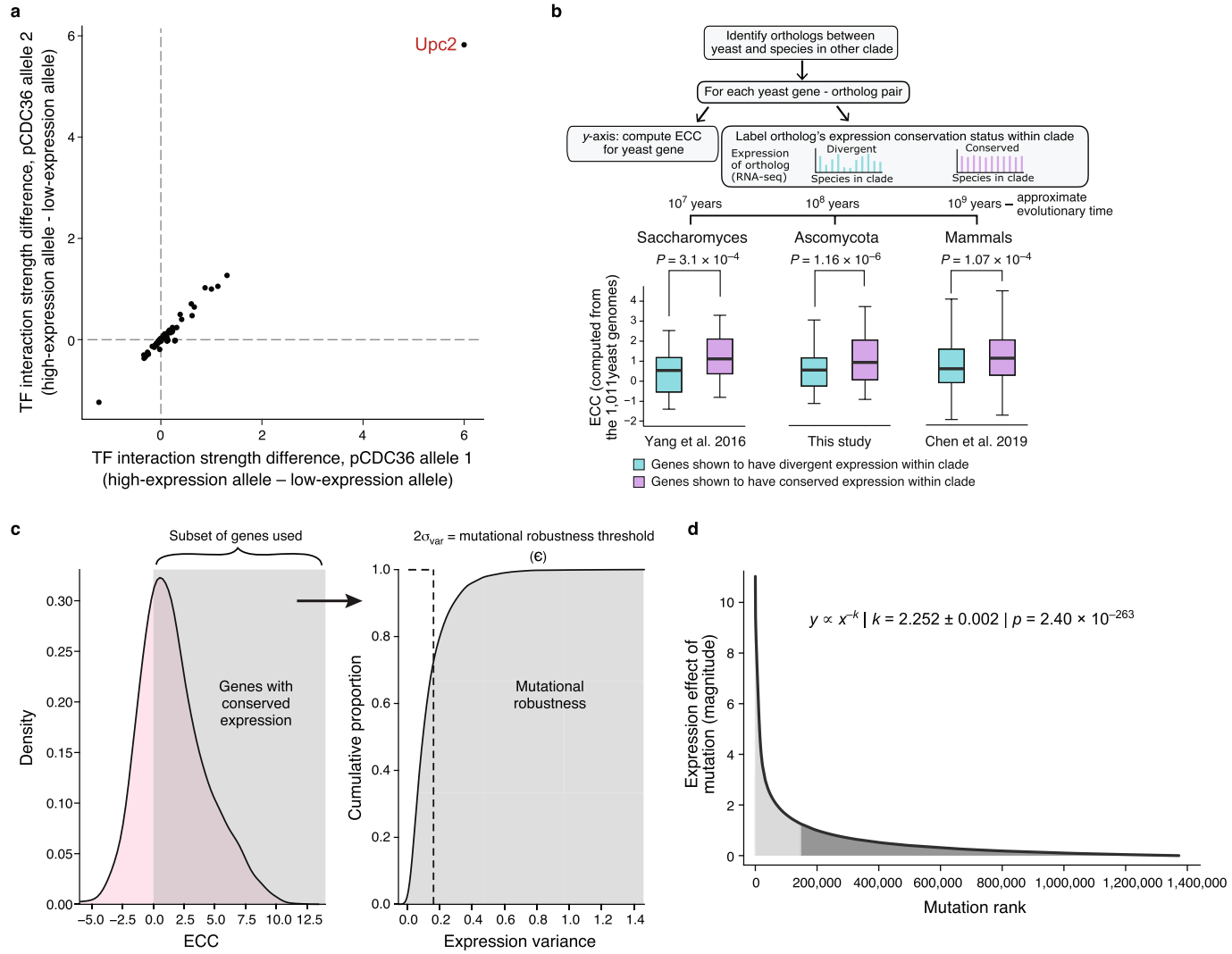
Found motif (DREME)	DREME *E*-value	Rank in native	Known factor	Known motif (YeTFaSCo)
	1.9×10^{-366}	1	Several (e.g., PHO4)	
	5.4×10^{-250}	5	REB1	
	1.0×10^{-184}	24	DOT6/TOD6	
	2.6×10^{-90}	Not found	TBF1	
	1.6×10^{-77}	7	Several (e.g. UPC2)	

Extended Data Fig. 2 | Characterization of sequence trajectories under strong competing selection pressures using the convolutional model. **a**, **b**, Expression is highly correlated between defined and complex medium. Measured (**a**) and predicted (**b**) expression in defined (*x* axis) and complex (*y* axis) medium for a set of test sequences measured in both media. Top left: Pearson's *r* and associated two-tailed *P* values. **c**, Opposing relationships between organismal fitness and *URA3* expression in two environments. Measured expression (*x* axis, using a YFP reporter) and fitness (*y* axis; when used as the promoter sequence for the *URA3* gene) for yeast with each of 11 promoters predicted to span a wide range of expression levels in complex medium with 5-FOA (red), where higher expression of *URA3* is toxic owing to *URA3*-mediated conversion of 5-FOA to 5-fluorouracil, and in defined medium lacking uracil (blue), where *URA3* is required for uracil synthesis. Error bars: Standard error of the mean (n = 3 replicate experiments). **d**–**f**, Competing expression objectives constrain adaptation. **d**, **e**, Difference in predicted expression (*y* axis) at each evolutionary time step (*x* axis) under selection to maximize (red) or minimize (blue) the difference between expression in defined and complex medium, starting with either native sequences (**d**, as Fig. 2h, n = 5,720) or random sequences (**e**, n = 10,000). **f**, Distribution of predicted expression (*y* axis) in complex (blue) and defined (red) medium at each evolutionary time step (*x* axis) for a starting set of random sequences (n = 10,000). Midline: median; boxes: interquartile range; whiskers: 5th and 95th percentile range. **g**, Motifs enriched within sequences evolved for competing objectives in different environments. Top five most enriched motifs, found using DREME[87] (Methods) within sequences computationally evolved from a starting set of random sequences to either maximize (left) or minimize (right) the difference in expression between defined and complex medium, along with DREME E-values, the corresponding rank of the same motif when using native sequences as a starting point, the probable cognate transcription factor and that transcription factor's known motif.

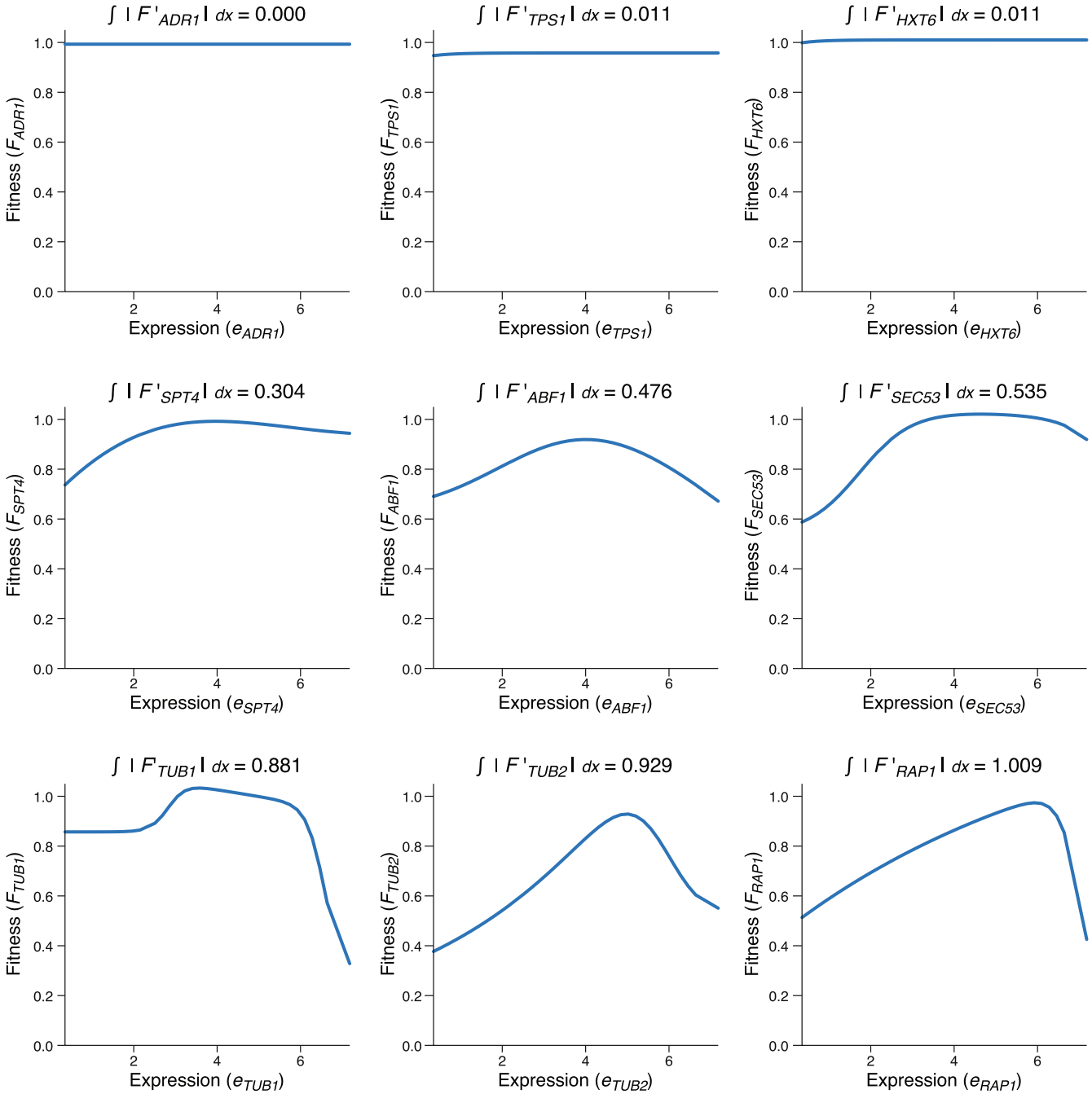


Extended Data Fig. 3 | See next page for caption.

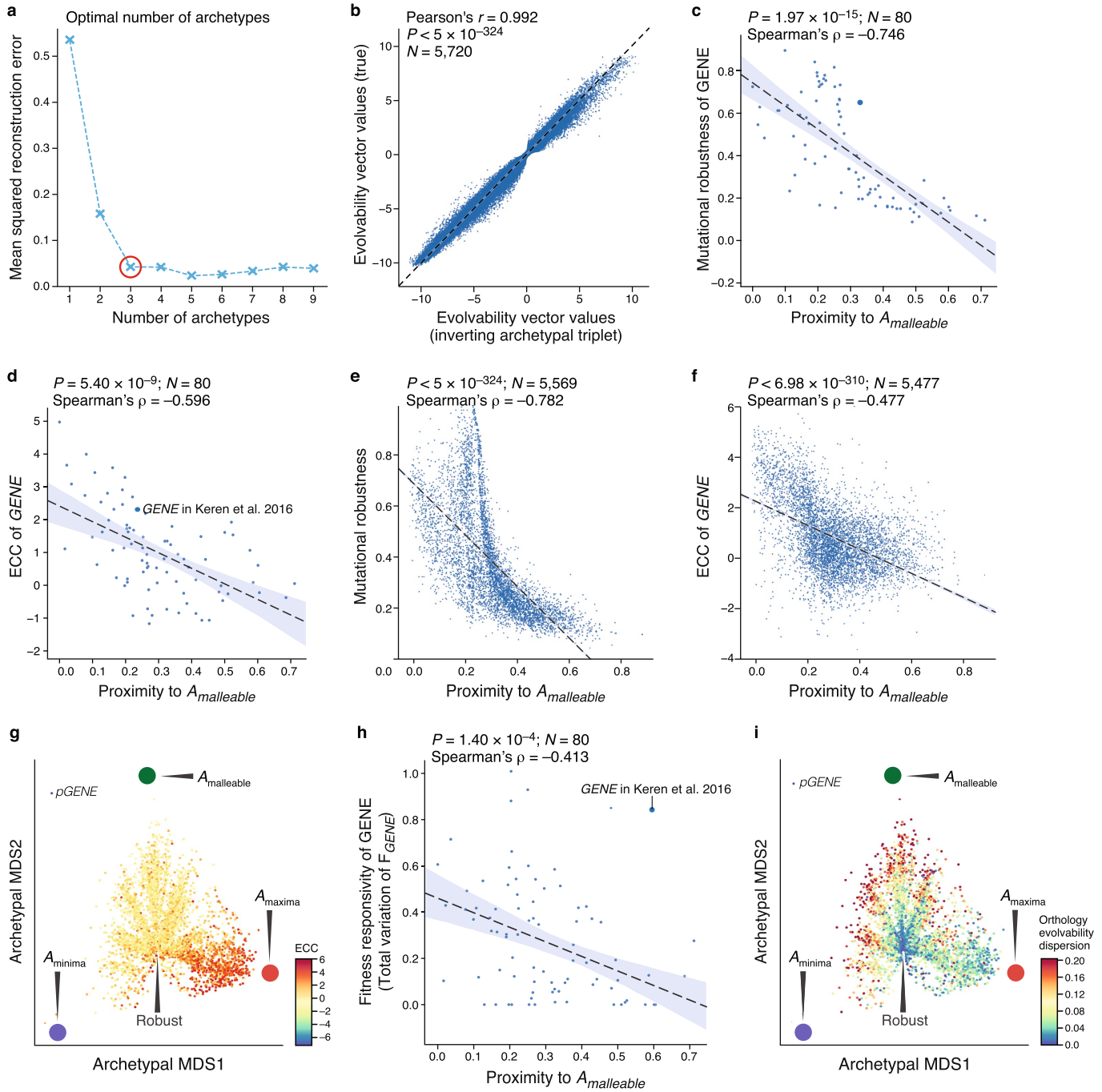
Extended Data Fig. 3 | The transformer sequence-to-expression model generalizes reliably and characterizes sequence trajectories under different evolutionary regimes. a–**d**, Prediction of expression from sequence in the complex (**a**, **b**) and defined (**c**, **d**) medium. Predicted (*x* axis) and experimentally measured (*y* axis) expression for (**a**, **c**) random test sequences (sampled separately from and not overlapping with the training data) and (**b**, **d**) native yeast promoter sequences containing random single base mutations. Top left: Pearson's *r* and associated two-tailed *P* value. Compression of predictions in the lower left results from binning differences during cell sorting in different experiments (Supplementary Notes). **e**, Predicted (*x* axis) and experimentally measured (*y* axis) expression in complex medium (YPD) for all native yeast promoter sequences. Pearson's *r* and associated two-tailed *P* values are shown. **f**, Predicted expression divergence under random genetic drift. Distribution of the change in predicted expression (*y* axis) for random starting sequences (n = 5,720) at each mutational step (*x* axis) for trajectories simulated under random genetic drift. Silver bar: differences in expression between unrelated sequences. **g**, **h**, Comparison of the distribution of measured (light grey) and transformer model predicted (dark grey) changes in expression (*y* axis) in complex medium (**g**, n = 2,983) and defined medium (**h**, n = 2,986) for synthesized randomly designed sequences at each mutational step (*x* axis). **i**, **j**, Predicted expression evolution under SSWM. Distribution of predicted expression levels (*y* axis) in complex medium (**i**, n = 10,322) and defined medium (**j**, n = 6,304) at each mutational step (*x* axis) for sequence trajectories under SSWM favouring high (red) or low (blue) expression, starting with 5,720 native promoter sequences. (**f**–**j**) Midline: median; boxes: interquartile range; whiskers: 5th and 95th percentile range. **k**–**r**, Comparison of model predicted expression for sequences synthesized previously for the random genetic drift and SSWM analyses. Experimentally measured (*y* axis) and transformer model predicted (*x* axis) expression level (**o**–**r**) or expression change from the starting sequence (**k**–**n**) in complex (**k**, **m**, **o**, **q**) or defined (**l**, **n**, **p**, **r**) medium using sequences from the random genetic drift (Fig. 2c, Extended Data Fig. 1e; **k**, **l**, **o**, **p** here) and SSWM (Fig. 2g, Extended Data Fig. 1g; **m**, **n**, **q**, **r** here) validation experiments. Top left: Pearson's *r* and associated two-tailed *P* values.



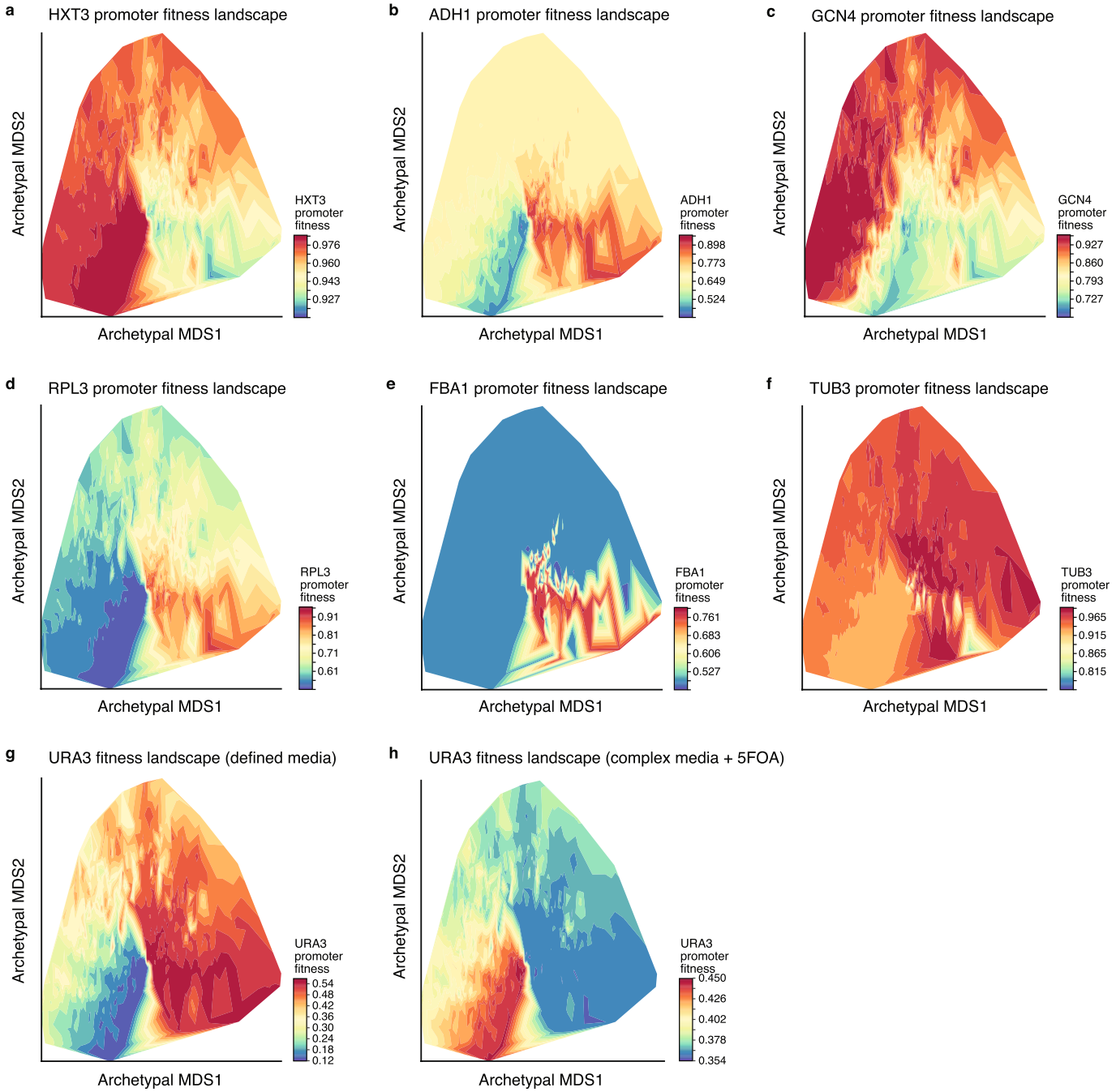
Extended Data Fig. 4 | Signatures of stabilizing selection on gene expression detected from regulatory DNA across natural populations. **a**, Expression-altering alleles in the CDC36 promoter are attributed primarily to altered UPC2 binding. Transcription factor interaction strength[26] (expression attributable to each transcription factor) difference between the high and low alleles (each point is a transcription factor) for each of two low expression alleles (allele 1: *x* axis; allele 2: *y* axis). Each low-expressing allele is compared to the high-expression allele with the most similar sequence (across all promoter sequences analysed from the 1,011 strains; $e_{\mathrm{TF},A_{high}} - e_{\mathrm{TF},A_{low}}$). **b**, Distribution of ECC (*y* axis, calculated from 1,011 *S. cerevisiae* genomes, top left) for *S. cerevisiae* genes whose orthologues have divergent (blue) or conserved (purple) expression (within *Saccharomyces* (left, n = 4,191), Ascomycota (middle, n = 4,910), or mammals (right, n = 199) (as determined by cross species RNA-seq, top right). *P* values: two-sided Wilcoxon rank-sum test. Midline: median; boxes: interquartile range; whiskers: 5th and 95th percentile range. **c**, Determination of expression change threshold for defining a 'tolerated mutation' to compute mutational robustness. We used all genes with an ECC consistent with stabilizing selection (ECC > 0; left), calculated the variance in predicted expression across the 1011 yeast strains for each gene, and chose the tolerable mutation threshold, ϵ, as two standard deviations of the distribution of the variance (right). ~73% of genes with ECC > 0 had an expression variation lower than ϵ. **d**, Distribution of the effects (magnitude; *y* axis) of mutations (rank ordered; *x* axis) on expression for all native regulatory sequences follows a power law with an exponent of 2.252. Shaded regions are equal in area.



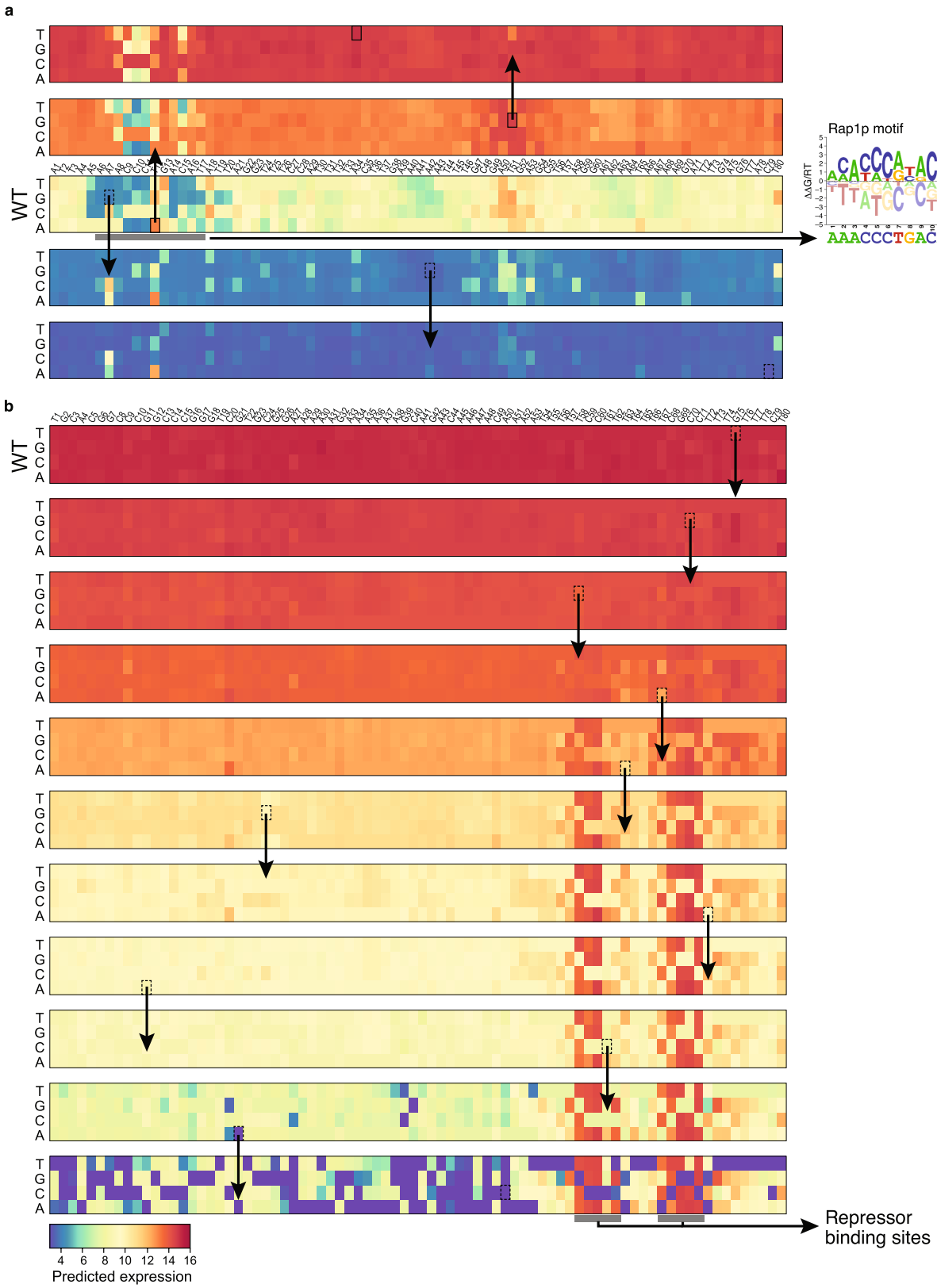
Extended Data Fig. 5 | Fitness responsivity of a gene as the total variation of its expression-to-fitness relationship F_{GENE} curves. Expression (*x* axis) and fitness (*y* axis) level curves for each select gene, fit from experimental measurements of expression and fitness across promoter variants by Keren et al[11]. Fitness responsivity calculated as the total variation in each curve is noted above each panel.



Extended Data Fig. 6 | Analysis of regulatory evolvability reveals sequence-encoded signatures of expression conservation from solitary sequences. a, Selection of optimal number of archetypes. Mean-square-reconstruction error (*y* axis) for reconstructing the evolvability vectors from the embeddings learned by the autoencoder for an increasing number of archetypes (*x* axis). Red circle: optimal number of archetypes selected as prescribed[45] by the 'elbow method'. **b**, The archetypal embeddings learned by the autoencoder accurately capture evolvability vectors. Original (*y* axis) and reconstructed (*x* axis) expression changes (the values in the evolvability vectors) for each native sequence (none seen by the autoencoder in training). Top left: Pearson's *r* and associated two-tailed *P* values. **c**–**f**, Evolvability space captures regulatory sequences' evolutionary properties. Proximity to the malleable archetype ($A_{malleable}$) (*x* axis) and mutational robustness (**c**, **e** *y* axis) or ECC (**d**, **f** *y* axis) for all yeast genes (**e**, **f**) or the gene for which fitness responsivity was quantified (**c**, **d**). Top right: Spearman's *ρ* and associated two-sided *P* value. 'L'-shape of relationship in **e** results from the robust cleft, A_{maxima}, and A_{minima} all being distal to $A_{malleable}$ (left side of plot). **g**, All native (S288C reference) promoter sequences (points) projected onto the archetypal evolvability space learned from random sequences; coloured by their ECC. Large coloured circles: evolvability archetypes. **h**, The proximity to the malleable archetype (*x* axis) and fitness responsivity (*y* axis) for the 80 genes with measured fitness responsivity. Top right: Spearman's *ρ* and associated two-tailed *P* values. Light blue error band: 95% confidence interval. **i**, All native (S288C reference) promoter sequences (points) projected on the evolvability space learned from random sequences; coloured by their mean pairwise distance in the archetypal evolvability space between all promoter alleles across the 1,011 yeast isolates for that gene (orthologue evolvability dispersion). Large coloured circles: evolvability archetypes.



Extended Data Fig. 7 | Visualizing promoter fitness landscapes in sequence space. Visualizing the fitness landscapes for the promoters of *HXT3* (**a**), *ADH1* (**b**), *GCN4* (**c**), *RPL3* (**d**), *FBA1* (**e**), *TUB3* (**f**), *URA3* (in defined medium) (**g**), *URA3* (in complex medium + 5FOA) (**h**). 1,000 promoter sequences represented by their evolvability vectors projected onto the 2D archetypal evolvability space and coloured by their associated fitness as reflected by their predicted growth rate relative to wild type (colour, Methods), estimated by first mapping sequences to expression with our model and then expression to fitness as measured and estimated previously[11].



Extended Data Fig. 8 | In silico mutagenesis of malleable and robust promoters. SSWM trajectories for (**a**) *DBP7*, a malleable promoter, and (**b**) *UTH1*, a robust promoter. Each subplot shows the *in silico* mutagenesis effects for how expression level (colour) changes when mutating each position (*x* axis) to each of the four bases (*y* axis) of each sequence (subplots) in the trajectories. The DNA sequence is indicated above each wild-type subplot (indicated with 'WT' at left). Arrows indicate the mutations selected at each step, which always correspond to the mutation of maximal effect; increasing expression goes up the figure from wild type and decreasing expression goes down. Part of the malleability of the *DBP7* promoter results from an intermediate-affinity Rap1p-binding site (grey bar). The first mutations in increasing- and decreasing-expression trajectories either increase or decrease (respectively) the affinity of this site. The *UTH1* promoter changes gradually in expression and evolves proximal repressor binding sites to dampen expression (grey bars).

nature research

Corresponding author(s): Eeshit Dhaval Vaishnav (edv@mit.edu), Carl G. de Boer (carl.deboer@ubc.ca), Aviv Regev (aviv.regev.sc@gmail.com)

Last updated by author(s): Dec 25, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specialized software was used for data collection.
Data analysis	All custom code used for the analysis is available on Github at https://github.com/1edv/evolution and https://github.com/de-Boer-Lab/CRM2.0. The code is also available in a directly executable form with all dependencies installed on CodeOcean at https://codeocean.com/capsule/8020974/tree. An interactive web application for our project is available at https://1edv.github.io/evolution/. Model architecture was written in Tensorflow v1.14 using Python v3.6.7. The list of open source libraries and functions used can be found at https://github.com/1edv/evolution/blob/master/aux.py and versions for all the libraries used can be found at https://github.com/1edv/evolution/blob/master/manuscript_code/evolution_env.yml. They are also listed below : - tensorflow 1.14 (https://www.tensorflow.org/) - matplotlib 3.0.3 (https://matplotlib.org/) - numpy 1.16 (https://numpy.org/) - scipy 1.4.1 (https://scipy.org/) - pandas 0.25 (https://pandas.pydata.org/) - biopython 1.74 (https://biopython.org/) - pydna 3.0.1 (https://pypi.org/project/pydna/) - scikit-learn 0.19 (https://scikit-learn.org/stable/) - seaborn 0.9 (https://seaborn.pydata.org/)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data generated for this study are available on NCBI's Gene Expression Omnibus, accession numbers GSE163045 and GSE163866. All processed data and trained models are available on Zenodo at https://zenodo.org/record/4436477.

Figures 1-4 and Extended Data Figures 1-8 all have associated source data.

All data is openly available.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not predetermined. We aimed to get as many promoters as possible for random libraries; more data would likely produce superior models, but the marginal gain is small since the models already perform sufficiently well for our purposes as shown by numerous validations.
Data exclusions	Specific promoters were excluded only if insufficient data had been collected to determine the expression level of that promoter. For example, if we saw no sequencing reads that matched to that promoter in the data, we excluded the promoter.
Replication	Similar conclusions were derived from multiple independently created datasets (by us and others) as described throughout the manuscript. Training datasets were done only in singlicate, but we compared the models created from one training dataset (defined media) or the other (complex media), and found that these were largely in agreement (Pearson's r = 0.966). For high-quality test datasets, we measured the same set of promoters in multiple media independently. These served as important measures of reproducibility (Pearson r = 0.978 between media). Analyses in Figure 1 and 2 were replicated with two independently created deep learning models which had similar results as noted in the text. The ECC (Figure 3) analyses were replicated with 5 models which were all found to have similar results.
Randomization	The promoter sequences used were randomly sampled from the sequence space through the process of random synthesis as described in the manuscript. Additionally, a random subset of yeast cells were transformed. The training, test and validation data were also randomized prior to learning and evaluating the models.
Blinding	The investigators were blinded to group allocation during data collection. The identities of the promoter sequences were unknown the investigators until after the experiments were performed and the measurements were made. The promoter sequences and expression levels were not blinded to the investigators during analysis because this would not be possible (we required knowing the promoter sequences and expression levels to train and test our models). Generally, we inspected data only in pre-defined groups (e.g. random sequences, trajectories, native sequences), rather than individually. Further, the scale of our data (tens of thousands to tens of millions of sequences) meant that it was not feasible to inspect individual sequences in any particular detail, limiting the opportunity for the kinds of bias that blinding could prevent.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	S288C (ATCC), Y8205 (Boone Lab, University of Toronto)
Authentication	Authentication based on nutritional requirements was performed (e.g. whether the yeast could grow in the presence or absence of certain nutrients). PCR of recombinant loci (URA3) was also performed.
Mycoplasma contamination	Yeast strains were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	The strains used in this study are not commonly misidentified lines.