RESEARCH ARTICLE SUMMARY

MOLECULAR BIOLOGY

Short tandem repeats bind transcription factors to tune eukaryotic gene expression

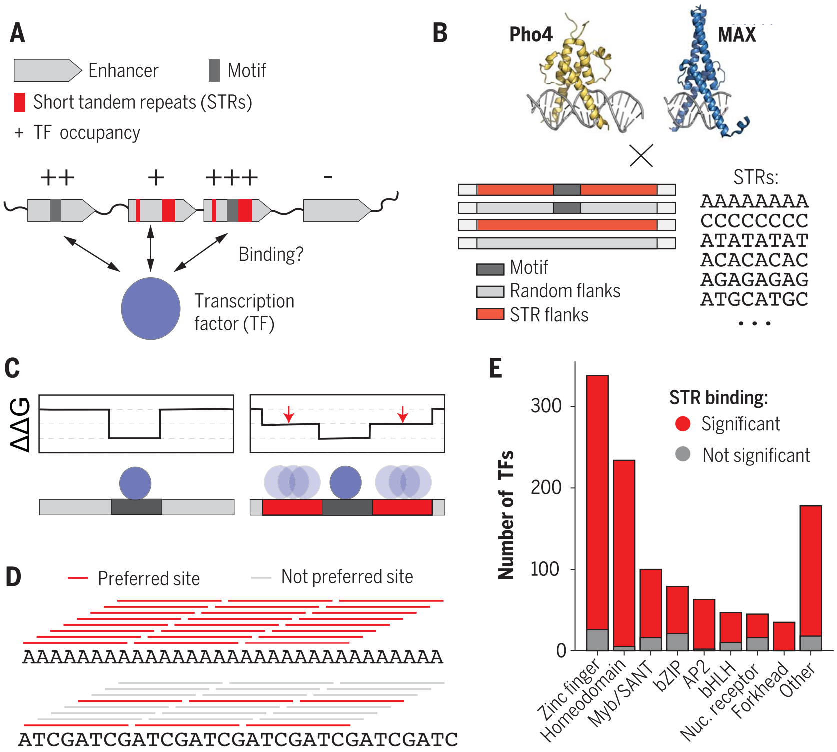
Connor A. Horton, Amr M. Alexandari†, Michael G. B. Hayes†, Emil Marklund†, Julia M. Schaepe†, Arjun K. Aditham, Nilay Shah, Peter H. Suzuki, Avanti Shrikumar, Ariel Afek, William J. Greenleaf, Raluca Gordân, Julia Zeitlinger, Anshul Kundaje, Polly M. Fordyce*

INTRODUCTION: Gene expression is regulated by transcription factor (TF) proteins that bind DNA-regulatory elements in the genome. Despite decades of research cataloging TF "motifs," these do not fully explain observed genomic binding in cells. Many TFs bind regions lacking motifs, whereas other regions with apparently strong motifs remain unoccupied, and emerging evidence suggests that the DNA sequence context surrounding motifs can strongly affect binding (see the figure, panel A). Short tandem repeats (STRs, consecutively repeated units of one to six nucleotides) provide a good example of these sequence contexts. STRs comprise ~5% of the human genome (compared with 1.5% for all protein-coding genes) and are enriched in enhancers. Variations in STR length have been associated with changes in gene expression and implicated in several complex phenotypes, such as schizophrenia, cancer, autism, and Crohn's disease. However, the mechanism by which STRs affect transcription remains unknown.

RATIONALE: One mechanism by which STRs could affect gene expression is by altering the affinity and/or kinetics of TF binding to regulatory DNA (see the figure, panel A). To investigate this, we used various high-throughput microfluidic binding assays (i.e., MITOMI, k-MITOMI, and STAMMP) and bioinformatic analyses to systematically quantify the impacts of different sequence contexts on TF binding. We measured affinities (K_ds) and kinetics (k_{off}s) for two basic helix-loop-helix TFs that bind a CACGTG E-box motif (Pho4 from *Saccharomyces cerevisiae* and MAX from *Homo sapiens*) to DNA sequences with or without an E-box motif surrounded by random sequence or multiple different types of STRs (see the figure, panel B).

RESULTS: Measured binding constants (K_ds) for 609 distinct TF-DNA combinations revealed that different STRs can alter binding affinities by >70-fold (see the figure, panel C), approaching or exceeding effects from mutating the consensus motif. Preferred STRs differed for Pho4 and MAX TFs, demonstrating that motifs are not sufficient to predict preferred STRs. Gel-shift assays and additional experiments using TF truncation constructs established that TFs directly bind STRs (see the figure, panel C) through their DNA-binding domains in the presence or absence of motifs. Although not predicted by standard mononucleotide models, the observed STR binding is well explained by a simple partition function model from statistical mechanics in which multiple repeated weak binding sites contribute additively to binding affinity (see the figure, panel D). Measured apparent dissociation rates (k_{off}s) for 106 TF-DNA combinations and kinetic modeling suggested that STRs primarily alter macroscopic apparent association rates and increase the local density of DNA-bound TFs. Finally, neural networks trained only on in vivo genome-wide chromatin immunoprecipitation data predict effects identical to those measured in vitro, suggesting that STR preferences play a substantial role in properly localizing TFs in cells.

CONCLUSION: Analysis of previously published protein-binding microarray and SELEX data suggests that ~90% of eukaryotic TFs preferentially bind at least one type of STR (see the figure, panel E). Because STRs are highly mutable, we propose that they should be considered an easily evolvable class of *cis*-regulatory elements. Preferred STRs need not resemble known motifs, suggesting a mechanism by which TF paralogs can be recruited to different regulatory regions and regulate distinct target genes. Although STRs maximize the number of potential weak binding sites, we anticipate that nonrepetitive sequence contexts containing many low-affinity binding sites should similarly increase binding. Thus, we propose that STRs function as "rheostats" to tune local TF concentration and binding responses to regulate gene expression in disease, development, and homeostasis. ■



STRs directly bind TFs to alter gene expression. (**A**) Schematic of enhancers, motifs, and STRs. (**B**) Schematic of TFs and DNA libraries tested in this study. (**C**) Favorable STRs alter energetic landscapes by directly binding TF DNA-binding domains. (**D**) Favorable STRs maximize potential preferred sites and contribute additively to binding energies. (**E**) STR binding by TFs is widespread.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: pfordyce@stanford.edu
†These authors contributed equally to this work.

Cite this article as C. A. Horton *et al.*, *Science* **381**, eadd1250 (2023). DOI: 10.1126/science.add1250

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.add1250

RESEARCH ARTICLE

MOLECULAR BIOLOGY

Short tandem repeats bind transcription factors to tune eukaryotic gene expression

Connor A. Horton[1]†, Amr M. Alexandari[2]‡, Michael G. B. Hayes[1]‡, Emil Marklund[1]‡, Julia M. Schaepe[3]‡, Arjun K. Aditham[3,4]§, Nilay Shah[5], Peter H. Suzuki[3], Avanti Shrikumar[2], Ariel Afek[6,7,8], William J. Greenleaf[1], Raluca Gordân[6,7,9,10], Julia Zeitlinger[5,11], Anshul Kundaje[1,2], Polly M. Fordyce[1,3,4,12]*

Short tandem repeats (STRs) are enriched in eukaryotic *cis*-regulatory elements and alter gene expression, yet how they regulate transcription remains unknown. We found that STRs modulate transcription factor (TF)–DNA affinities and apparent on-rates by about 70-fold by directly binding TF DNA-binding domains, with energetic impacts exceeding many consensus motif mutations. STRs maximize the number of weakly preferred microstates near target sites, thereby increasing TF density, with impacts well predicted by statistical mechanics. Confirming that STRs also affect TF binding in cells, neural networks trained only on in vivo occupancies predicted effects identical to those observed in vitro. Approximately 90% of TFs preferentially bound STRs that need not resemble known motifs, providing a *cis*-regulatory mechanism to target TFs to genomic sites.

Short tandem repeats (STRs), consisting of 1- to 6-base pair (bp) units repeated consecutively, comprise 5% of the human genome (figs. S1 to S3), compared with 1.5% for protein-coding genes (*1*, *2*), with a median STR length of 29 bp (fig. S4). STRs are enriched in *cis*-regulatory elements across eukaryotic genomes (*3*), including in humans [~25% of enhancers contain an STR (*4*, *5*); fig. S5], and can activate or repress transcription in *Homo sapiens* (*5*–*21*), *Mus musculus* (*22*, *23*), *Saccharomyces cerevisiae* (*3*), *Drosophila melanogaster* (*24*, *25*), and others (*26*). Dinucleotide STRs are associated with broad activity of *cis*-regulatory elements across cell types in *D. melanogaster* (*27*), and variation in STRs has been proposed to account for "missing heritability" in genome-wide association studies (*5*, *28*). Finally, population-level genomic studies have linked noncoding STR polymorphisms to autism (*29*, *30*), schizophrenia (*31*), height (*31*), and Crohn's disease (*5*).

Despite their widespread prevalence and documented effects on gene expression, the physical mechanism by which STRs affect transcription remains unclear. STRs have been proposed to modulate transcription by changing the intrinsic affinity of histone proteins for DNA, thereby changing nucleosome occupancy (*3*, *22*, *24*, *32*). However, STRs have not been shown to directly alter chromatin accessibility other than the example of nucleosome-disfavoring poly(A) tracts (*33*). Alternatively, polymorphisms in STR length could alter distances between multiple motifs or between motifs and core promoter elements, disrupting regulatory grammar (*34*–*36*). However, genome-wide studies suggest that the syntax of cooperative transcription factor (TF) interactions at enhancers is unlikely to be perturbed by changes in motif spacing (*37*–*39*). Theoretical work has suggested that "sequence symmetries" (i.e., repetitiveness) alone contribute to nonspecific TF binding, with maximum effects for homopolymer sequences (*40*, *41*), and in vitro binding measurements and bioinformatic analyses have suggested that STRs affect TF-DNA binding in the absence of specific base pair recognition (*40*, *42*–*46*). Nevertheless, prevailing models of TF specificity do not predict observed specific binding to STRs (*42*), and the magnitudes of their energetic effects and potential impacts on binding kinetics remain unexplored. Here, we used multiple high-throughput microfluidic binding assays [MITOMI (*47*, *48*), *k*-MITOMI (*49*), and STAMMP (*50*)] to systematically investigate how STRs influence equilibrium binding and kinetics for two different basic helix-loop-helix (bHLH) TFs.

Results

Quantitative measurements establish STRs alter TF binding affinities

The bHLH TFs Pho4 [a yeast TF involved in the phosphate starvation response (*51*, *52*)] and MAX [a human TF involved in cell proliferation, differentiation, and apoptosis (*53*, *54*)] each bind an E-box regulatory element (Fig. 1A and table S1). To test the impact of STRs on binding, we quantified the binding of each TF to 17 DNA sequences containing either a moderate-affinity extended E-box sequence (GTCACGTGAC) or a random sequence ("no motif") flanked by 13 bp of either random sequence or STRs previously shown to enhance binding (*42*) (Library 1; Fig. 1B and table S2) through MITOMI microfluidic binding assays (Fig. 1C, figs. S6 to S9, and table S3). Measured binding for each DNA sequence over multiple concentrations can be combined with calibration curves (figs. S10 and S11 and table S4) to extract the dissociation constant (K_d) by quantifying concentration-dependent TF binding and globally fitting Langmuir isotherms (Fig. 1C; see the materials and methods).

Measured Library 1 $\Delta\Delta G$s spanned ~2.6 and 3.1 kcal/mol with a mean root mean squared error (RMSE) between replicates of ~0.53 and 0.31 kcal/mol for Pho4 and MAX, respectively [Fig. 1D, figs. S12 to S21; see additional data at (*55*)]. DNA sequences with a motif surrounded by STRs were consistently bound 0.23 to 0.90 kcal/mol tighter than those with a motif surrounded by random sequences, corresponding to an ~1.5- to 4.6-fold change in predicted affinity (Fig. 1, D and E). Distributions of measured $\Delta\Delta G$s for sequences containing STRs were statistically significantly different from those with random sequences (as assessed by bootstrap hypothesis testing with a Bonferroni-corrected significance threshold; fig. S22 and table S5), and these effects scaled with STR length (fig. S23). Measured $\Delta\Delta G$s did not change with ~5-fold differences in protein concentration, confirming that DNA was in vast excess of available protein (figs. S24 and S25). Measured $\Delta\Delta G$s were also consistent when using either wheat germ extract or Tris-buffered saline (TBS) as a binding buffer (fig. S26), and negative control experiments assessing binding to enhanced green fluorescent protein (eGFP) alone showed no variability above the background RMSE (maximum deviation of ±0.5 kcal/mol; fig. S27). Linear mononucleotide specificity models such as the position-specific affinity matrix (PSAM) predicted a <0.1 kcal/mol effect for all flanking sequences but one ("Motif + GT/AC repeat 2") (fig. S28), establishing that the measured effects are not due to cryptic consensus sites in flanking sequences.

[1] Department of Genetics, Stanford University, Stanford, CA 94305, USA. [2] Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [3] Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [4] ChEM-H Institute, Stanford University, Stanford, CA 94305, USA. [5] Stowers Institute for Medical Research, Kansas City, MO 64110, USA. [6] Center for Genomic and Computational Biology, Duke University School of Medicine, Durham, NC 27710, USA. [7] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA. [8] Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 7610001, Israel. [9] Department of Computer Science, Duke University, Durham, NC 27708, USA. [10] Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27710, USA. [11] The University of Kansas Medical Center, Kansas City, KS 66103, USA. [12] Chan Zuckerberg Biohub, San Francisco, CA 94110, USA.
***Corresponding author. Email: pfordyce@stanford.edu**
†Present address: Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA.
‡These authors contributed equally to this work.
§Present address: Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA.

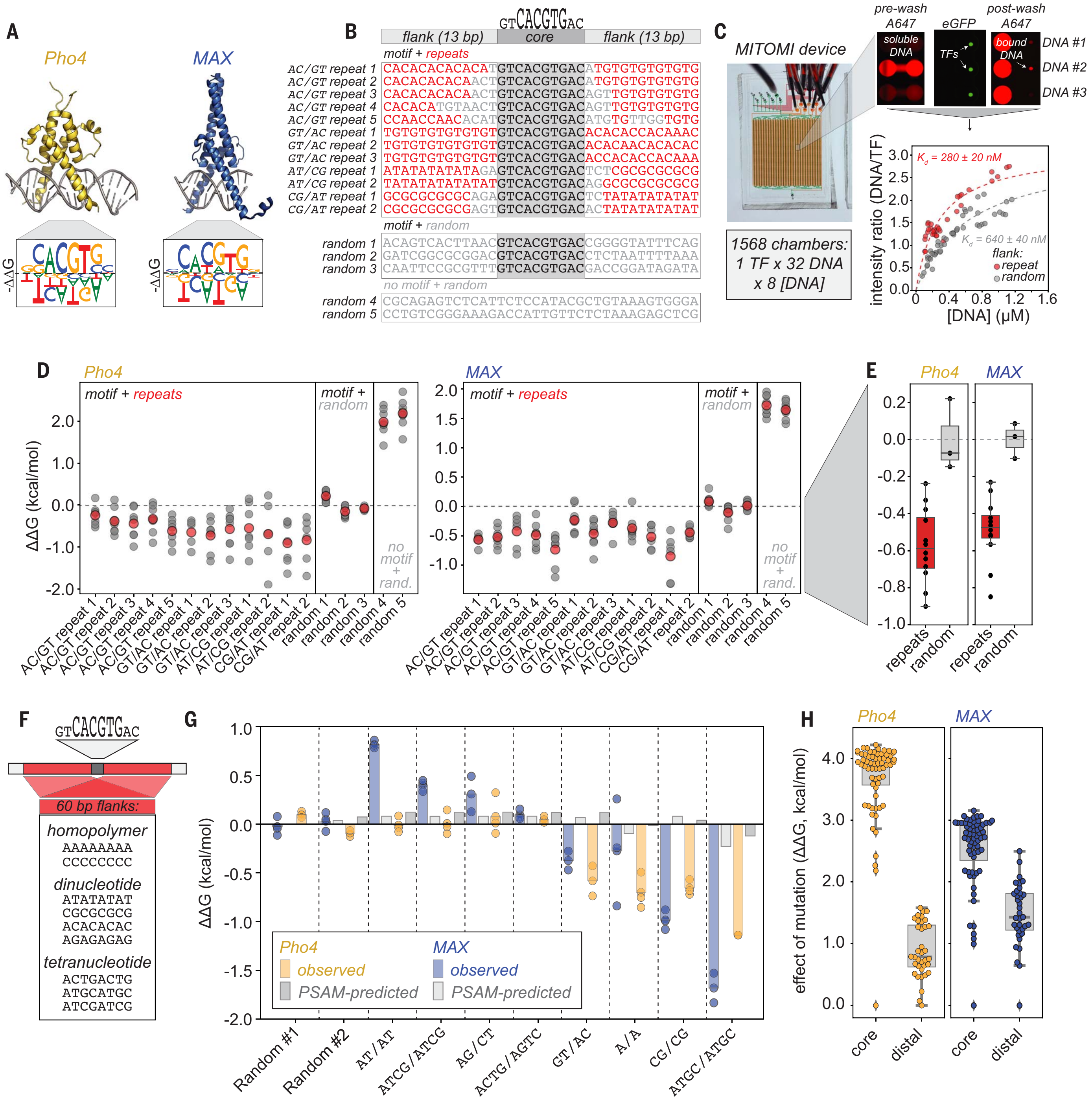


Fig. 1. Repetitive flanking sequences alter TF-DNA binding affinities in a sequence-specific manner. (**A**) Crystal structures and PSAM logos (47) for Pho4 [Protein Data Bank (PDB) ID: 1a0a] and MAX (PDB ID: 1hlo). (**B**) Library 1: 17 DNA sequences with either an extended (10-bp) E-box motif (dark gray) or random (light gray) sequence surrounded by 13 bp on either side of repetitive (red) or random (light gray) sequence. (**C**) MITOMI microfluidic device (left) and zoomed-in view of three chambers (top right) showing solubilized DNA during incubation (prewash A647), immobilized TFs (eGFP), and TF-bound DNA after washing (postwash A647). Bottom right shows representative concentration-dependent binding for DNA sequences containing an extended E-box surrounded by either repetitive (red) or random (gray) flanks. (**D**) Measured ΔΔG values across all Library 1 sequences for Pho4 (left) and MAX (right). ΔΔGs were calculated relative to the overall median value for oligonucleotides bearing an E-box consensus surrounded by random flanking sequence. Light gray dots show all measurements; red circles indicate median values per oligo. (**E**) Median values (black markers and box plots) for all sequences containing either repetitive (red) or random (gray) flanking sequences for Pho4 (left) and MAX (right). (**F**) Library 2: 10 DNA sequences containing a central extended (10-bp) E-box motif surrounded by 60 bp on either side of listed homopolymeric, dinucleotide, or tetranucleotide repeats. (**G**) Measured ΔΔG values across all Library 2 sequences for Pho4 (gold) and MAX (blue). ΔΔGs were again calculated relative to the overall median value for oligonucleotides bearing an E-box consensus surrounded by random flanking sequence. Gray bars indicate magnitude of effects predicted by PSAMs. (**H**) Observed effects on ΔΔG for mutating single nucleotides within the CACGTG core E-box (core) (47) versus altering flanking sequence within Library 2 (distal) for Pho4 (left, gold) and MAX (right, blue) overlaid on boxplots (gray).

Magnitude of STR effects on affinity depends on STR sequence

To determine how STR sequence alters binding, we designed a DNA library containing either an extended consensus E-box motif or a random sequence surrounded on each side by 60-bp flanks (the approximate mean length of STRs in humans and budding yeast; fig. S4) composed of random sequence or homopolymer, dinucleotide, or tetranucleotide STRs (Library 2; Fig. 1F and table S6; CG/AT indicates a CG repeat on one side of the motif and an AT repeat on the other). Because repetitive sequence extension can be technically challenging, we visualized extension through denaturing gel electrophoresis (fig. S29) and quantified binding affinities only for sequences that extended successfully (Fig. 1G, figs. S30 to S36, and table S6). The observed effects ranged from increasing affinity by 1.7 kcal/mol (18-fold) to reducing affinity by 0.8 kcal/mol (4-fold). Whereas ATGC STRs enhanced binding for both Pho4 and MAX, other STRs (AT/AT, ATCG/ATCG, and AG/CT) were deleterious for MAX only (Fig. 1G, figs. S37 and S38, and table S7). As for Library 1, the distribution of measured ΔΔGs for multiple sequences with STRs differed significantly from those with different random sequences (fig. S39 and table S8), results did not change with surface protein density (figs. S40 and S41), no sequence-specific binding was detected for an eGFP-only negative control (fig. S42), and the observed effects were inconsistent with PSAM-based models of specificity (Fig. 1G). Effects also diverged significantly for Pho4 and MAX (Fig. 1G), signifying that "consensus" binding motifs are insufficient to predict STR preferences. Energetic contributions of flanking sequences approached or exceeded those associated with mutating core consensus residues ([47]), particularly for MAX, suggesting that STRs could play a significant role in proper TF localization in vivo (Fig. 1H).

STRs alter affinities by directly binding TFs

The observed STR effects suggest two possible mechanistic models (Fig. 2A). In the first, STRs could enhance TF binding to the core consensus site, perhaps by altering local DNA "shape" ([56–60]) (Fig. 2A, top). This model predicts that STRs should only alter binding in the presence of a core motif and that TF-DNA stoichiometry should not depend on flanking sequence. The second model is that STRs could represent additional binding sites (Fig. 2A, bottom). This model predicts that STRs should enhance binding regardless of whether they flank a consensus motif and that multiple TFs will bind oligonucleotides containing STRs.

Concentration-dependent binding for Pho4 and MAX was clearly stronger for sequences containing favorable STRs even in the absence of a motif (Fig. 2, B and C). Moreover, energetic effects of STRs did not correlate with predicted DNA shape parameters (figs. S43 and S44), and circular dichroism spectroscopy ruled out enhanced binding resulting from STR-dependent structural transitions between B- and Z-form DNA (fig. S45). Finally, electrophoretic mobility shift assays (EMSAs) using Alexa Fluor 647–labeled double-stranded DNA (dsDNA) and increasing concentrations of eGFP-tagged (Fig. 2D and fig. S46) or untagged MAX (fig. S47) revealed supershifted bands at higher MAX concentrations for DNA sequences containing STRs, consistent with multiple TFs binding a single DNA molecule. Together, these experiments demonstrate that STRs modulate TF-DNA affinity by directly binding TFs in vitro.

Statistical mechanical models integrating data across experimental platforms accurately predict STR effects

Universal protein-binding microarray (uPBM) experiments measure binding of fluorescently tagged TFs to surface-immobilized DNA duplexes containing all possible 8-mer DNA sequences, providing comprehensive measurements of TF-DNA specificity in an alternate (flipped) experimental configuration relative to MITOMI ([61–64]). To determine whether previously published uPBM measurements also reveal enhanced binding of Pho4 and MAX to specific STRs, we calculated the median intensity for all probes containing each of the 65,538 possible DNA 8-mers and then calculated a Z score for each 8-mer relative to this distribution (Fig. 2, E and F). As expected, probes containing 8-mer variants of the known E-box CACGTG consensus were bound very strongly by Pho4 and MAX, with Z scores of 40 to 80 (Fig. 2F). Consistent with MITOMI results, favorable repeats were bound statistically significantly above background after Bonferroni correction for both MAX (ATGC, $Z = 15.1, P = 4 \times 10^{-127}$; CG, $Z = 8.3, P = 5 \times 10^{-40}$; and AC, $Z = 5.0, P = 1 \times 10^{-15}$) and Pho4 (ATGC, $Z = 10.7, P = 7 \times 10^{-72}$; GC, $Z = 3.9, P = 3 \times 10^{-11}$; and AC, $Z = 5.4, P = 9 \times 10^{-20}$; Fig. 2F).

Next, we combined information from the PBM and MITOMI experiments to determine whether partition function models from statistical mechanics improve binding predictions by accurately accounting for flanking sequence effects (Fig. 2, E and G). MITOMI-measured ΔΔGs and log-transformed gcPBM intensities for MAX binding to 32 probes (figs. S48 to S51 and tables S9 and S10) were strongly anticorrelated ($R_p^2 = 0.89$) over a wide dynamic range (2.5 kcal/mol) (fig. S52), confirming reports that PBM intensities can report on affinities ([62], [65–68]). This allowed us to compute a partition function from intensities and predict Pho4 and MAX ΔΔGs for all DNA Library 1 and 2 sequences (Fig. 2G; see the materials and methods). For DNA Library 1, which contains intact, mutated, or ablated E-box consensus sequences surrounded by 13-bp variable flanking sequences, partition function–based predictions significantly improve agreement with measured ΔΔGs over standard PSAM predictions (R_p^2 = 0.91 versus 0.66 and R_p^2 = 0.93 versus 0.74 for Pho4 and MAX, respectively) (fig. S53). For DNA Library 2, in which all sequences contain an E-box but differences in flanking sequences can change measured ΔΔGs by up to 1.6 and 2.5 kcal/mol for Pho4 and MAX, partition function–based calculations were substantially better correlated with measurements than PSAM models (R_p^2 = 0.71 versus 0.09 and R_p^2 = 0.81 versus 0.02 for Pho4 and MAX, respectively) (Fig. 2G and fig. S54). Returned fit parameters from these linear regressions allow calibration of partition function–based predictions in energetic space with as few as nine thermodynamic measurements (K_ds or ΔΔGs; fig. S55; see the materials and methods).

To determine whether sequencing-based selection experiments also reveal binding for MAX to the same STRs, we quantified the frequency with which each 8-mer DNA sequence appeared within the TF-bound fractions in the SMiLE-seq ([69]) and SELEX-seq ([70]) datasets, converted frequencies to Z scores, and again used these Z scores in a partition function to predict binding to DNA Library 1 and 2 sequences. Predicted binding was well correlated with observations for both libraries (R_p^2 = 0.86, 0.75 for Library 1 and R_p^2 = 0.59, 0.81 for Library 2 for SMiLE-seq and SELEX-seq, respectively; figs. S56 and S57).

Even weakly preferred STRs enhance binding by increasing the number of preferred microstates

Preferred repeats for Pho4 and MAX (e.g., CG and ATGC) do not resemble the known E-box consensus, as evidenced by a failure of PSAM-based models to predict the observed effects (Figs. 1H and 2G and figs. S28 and S53). Why, then, do repeats recruit TFs? By virtue of being repetitive, STRs create multiple identical binding sites that are equally probable binding microstates (Fig. 2H), and STRs (in particular, homopolymers) maximize binding entropy and therefore minimize Gibbs free binding energy when enthalpy is kept constant (see the supplementary text). To estimate the energetic magnitude of this statistical effect, we conducted Monte Carlo simulations that randomly sample from observed energy distributions to mimic either random or homopolymeric sequences (fig. S58; see the materials and methods). These simulations revealed that increasing repetitiveness alone can contribute up to 0.3 kcal/mol mean binding energy through entropic effects for sequences <60 bp (fig. S58). However, effects are considerably stronger for STRs with affinities only slightly above background binding: 57-bp dinucleotide STRs with

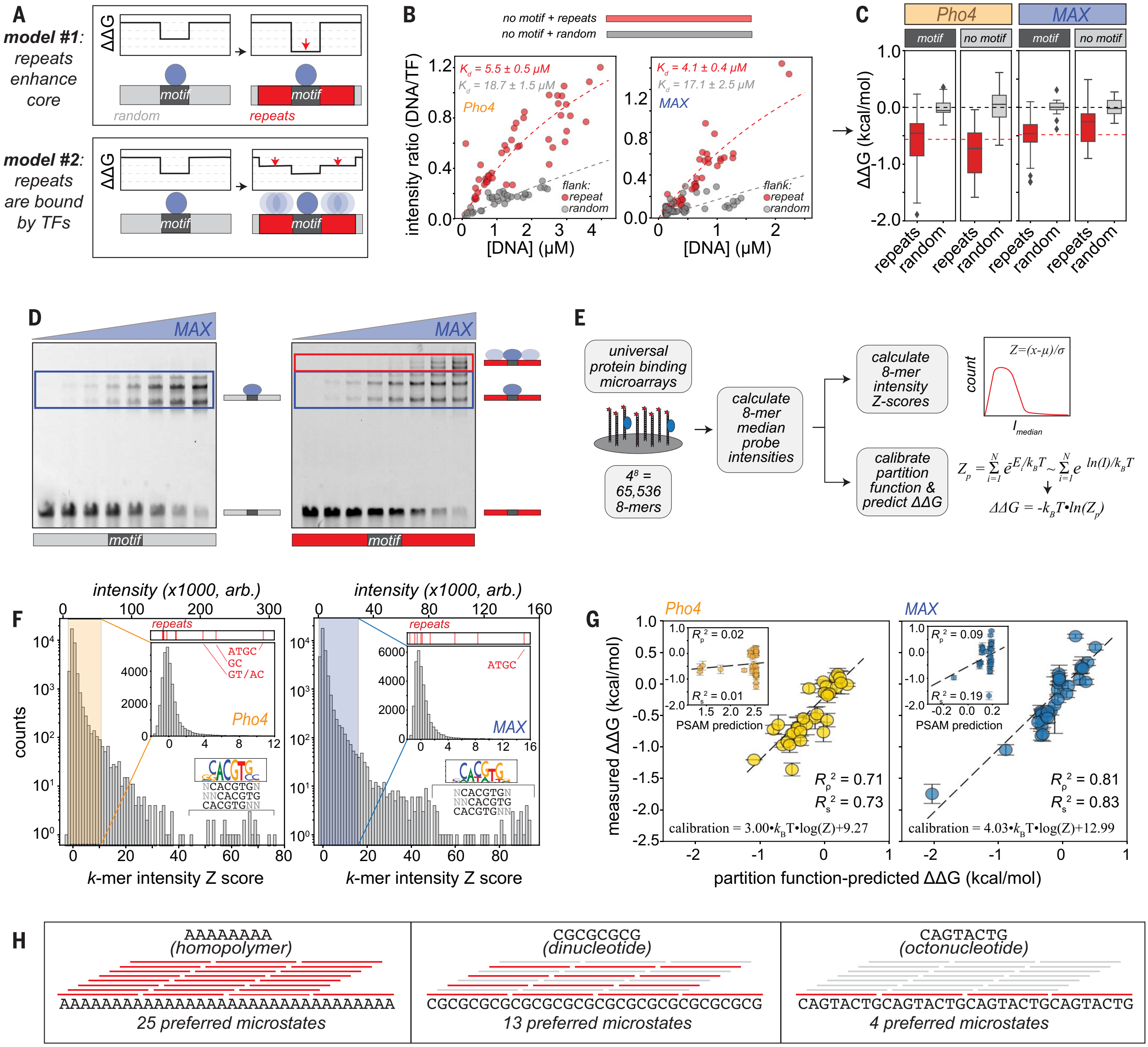


Fig. 2. STRs are directly bound by TFs with observed affinities that can be accurately predicted by statistical mechanics. (**A**) Models explaining how repetitive flanking sequences could enhance TF binding affinities. (**B**) Representative concentration-dependent binding for Pho4 (left) and MAX (right) interacting with DNA sequences containing either repetitive (red) or random (gray) sequences in the absence of an E-box motif. (**C**) Box plots of relative binding energies (ΔΔGs) for Pho4 and MAX binding to oligonucleotides with repetitive (red) or random (gray) sequence flanking an extended E-box consensus (dark gray) or random sequence (light gray); black and red dashed lines indicate median overall affinities. (**D**) EMSAs for increasing concentrations of eGFP-tagged MAX interacting with Alexa Fluor 647–labeled dsDNA duplexes containing a central extended E-box surrounded by random (left) or repetitive (right) sequences. Blue boxes highlight TF complexes bound to the core motif; red boxes highlight supershifted species with additional bound TFs. Native gel electrophoresis reveals MAX alone runs as three bands, likely representing MAX homodimers, MAX monomers, and eGFP-only truncation constructs (fig. S46). (**E**) Pipeline for calculating 8-mer intensity *Z* scores from universal PBM data and calibrating partition function scores to predict binding (see the materials and methods). (**F**) Log-linear histograms of intensity *Z* scores for all 8-mers for Pho4 (left) and MAX (right). Inset shows linear-linear plots that highlight background binding distributions and *Z* scores of the STRs measured in this study (red bars, top). (**G**) Scatter plots, linear regressions, and correlation coefficients for measured ΔΔGs versus calibrated partition function–predicted scores across all measured repeats for Pho4 (left) and MAX (right). (**H**) Schematic showing possible microstates as a function of sequence.

intensity *Z* scores of 1 to 2 or 5 to 10 are predicted to enhance binding by 0.6 and 1.4 kcal/mol (10-fold), respectively (fig. S58). Further validating these effects, partition function–predicted energy distributions for Pho4 and MAX binding 10,000 simulated sequences containing E-box consensus sites flanked by either random sequence or STRs showed that the most-favorable STRs bound more strongly than all 10,000 random sequences (fig. S59), a result not predicted by analogous simulations using mononucleotide models (fig. S60). A partition function model is purely additive, and additional mechanisms of cooperativity [e.g., allostery, avidity, and allovalency (71)] are not necessary to explain effects of STRs on in vitro binding.

STRs are directly bound by TF DNA-binding domains

Our results thus far had established that TFs directly bind STRs but did not identify which portion of the TF recognizes them. STRs may be recognized by intrinsically disordered regions (IDRs) outside of TF DNA-binding domains (DBDs) (72) or by DBDs themselves. To distinguish between these, we compared binding for eGFP-tagged full-length Pho4, the DBD alone, or the non-DBD alone to six DNA sequences containing either the extended E-box motif (GTCACGTGAC) or no motif surrounded by random sequence or favorable or moderately favorable STRs (ATGC and CG/AT, respectively; Fig. 3, A to C; figs. S61 to S63; and tables S11 and S12). Both full-length and DBD-only constructs showed enhanced binding to repeats (Fig. 3, A to C, and figs. S61 and S62) with strongly correlated measured K_d values ($R_p{}^2$ = 0.99; Fig. 3C). Consistent with prior reports that IDRs outside of the DBD can inhibit DNA binding (73–77), fluorescence intensity ratios (DNA bound per surface-immobilized TF) were consistently lower for the full-length construct (Fig. 3B). By contrast, the Pho4 non-DBD did not bind DNA with either random or the most-favorable ATGC STR flanking sequences above background levels, and measured K_ds were uncorrelated with the full-length construct (Fig. 3C and fig. S63). Although the Pho4 non-DBD exhibited detectable binding to the moderately favorable CG/AT STR (fig. S63), binding was extremely weak (K_d > 15 μM) and disappeared in the absence of a CACGTG motif (fig. S63), inconsistent with observations for full-length Pho4 (Fig. 2, B and C).

Because MAX has an extremely small non-DBD (49 residues versus 202 residues for Pho4; fig. S64A), we anticipated the MAX non-DBD was unlikely to bind STRs. To test this, we compared 8-mer Z scores between previously published uPBM (62) and SELEX-seq (70) data for full-length MAX and the DBD alone (fig. S64). If the MAX non-DBD binds STRs, then we would expect the full-length construct to return higher Z scores for favorable STRs compared with the DBD alone. Instead, all 8-mer Z scores were linearly correlated between constructs (R^2 = 0.72 and 0.90 for uPBM and SELEX-seq, respectively). Together, these analyses demonstrate that Pho4 and MAX recognize STRs through their DBDs.

To investigate which residues within the Pho4 DBD mediate STR recognition, we used STAMMP (50) (Fig. 3, D and E) to recombinantly express and purify 221 Pho4 variants containing systematic amino acid mutations within and surrounding the DBD (Fig. 3F and table S13) and quantify concentration-dependent binding for each variant interacting with DNA sequences containing a motif flanked by either random sequence or favorable CG dinucleotide STRs (Fig. 1G). Across nine STAMMP experiments, 214 of 221 variants showed strong expression (Fig. 3E, figs. S65 and S66, and table S14), and concentration-dependent binding was well fit by a Langmuir isotherm across both DNA sequences (Fig. 3G and figs. S67 to S72), yielding 6139 individual TF-DNA K_d measurements. After normalization between experiments, measured energetic effects were consistent across experiments (<0.48 kcal/mol RMSE) and spanned >4 kcal/mol (figs. S69 and S72).

We then compared measured ΔΔGs for each mutant relative to the wild-type (WT) TF across DNA sequences, reasoning that residues involved in STR recognition should differentially affect affinity upon mutation (Fig. 3H). Nearly all mutants altered binding affinities equally across DNA sequences, but E259D showed significantly enhanced binding to CG dinucleotide–flanking sequences (Fig. 3H and figs. S73 and S74; Z score of residual = 6.0, P = 1.7 × 10^{-9}, ΔΔΔG ≈ 0.73 kcal/mol). In the Pho4 crystal structure, E259 directly contacts nucleotides from both strands at the CACGTG position (78) (Fig. 3I), and comparisons of measured affinities for WT Pho4 and E259D revealed that although the WT Pho4 showed a strong preference for the canonical E-box motif (CACGTG), E259D showed equal, weak (100-fold lower) binding to the canonical E-box and a motif mutated at this position (CACGCG) (50) (Fig. 3J). These observations are consistent with a model in which increased promiscuity of the E259D binding energy landscape leads to an effective increase in preference for CG dinucleotide repeats (Fig. 3K).

STRs increase apparent macroscopic association rates

To investigate how flanking sequences alter TF binding kinetics, we leveraged k-MITOMI (49) (Fig. 4A) to quantify dissociation rates for Pho4 and MAX interacting with DNA sequences containing an extended E-box motif (GTCACGTGAC) surrounded by 60-bp flanks composed of random sequence or eight different STRs that extended properly (homopolymer: A/A; dinucleotide: AT/AT, AG/CT, GT/AC; tetranucleotide: ACGT/ACGT, ATCG/ATCG, ACTG/AGTC, ATGC/ATGC). Specifically, we iteratively (i) closed valves to trap TF-bound DNA, (ii) introduced a high-affinity unlabeled DNA competitor, (iii) opened valves for 1 to 4 s to allow fluorescently labeled DNA to dissociate, (iv) closed valves and washed out unbound material, and (v) imaged all device chambers (Fig. 4A). Excess unlabeled DNA competitor outcompeted rebinding to ensure accurate rate measurements. Decreases in measured Alexa Fluor 647/eGFP (DNA/TF) intensity ratios over time were well fit by a single exponential for Pho4 and MAX [Fig. 4B and figs. S75 and S76; see additional data at (55)]; rates typically varied by <3-fold across experiments before normalization (figs. S77 and S78). For both Pho4 and MAX, different 60-bp flanking STRs changed apparent rates of dissociation ($k_{\text{off,apparent}}$) from the entire dsDNA sequence only slightly (<1.7-fold, less than noise between experiments) (Fig. 4C and figs. S77 and S78). By contrast, inferred apparent on rates ($k_{\text{on,apparent}} = k_{\text{off,apparent}}/K_d$, calculated assuming a two-state model in which DNA is either bound or unbound) were substantially altered (Fig. 4C and figs. S79 to S83). These results were consistent across different normalization schemes (figs. S84 to S88), with favorable STRs increasing macroscopic apparent on-rates by 7- to 54-fold for Pho4 and MAX, respectively, suggesting that the observed changes in affinity were primarily caused by altered macroscopic apparent association rates (Fig. 4C and fig. S87).

STRs increase the density of weakly bound TFs near target motifs

STRs are enriched near binding sites of stress-response TFs in budding yeast that likely require a rapid transcriptional response (3), suggesting that STRs could reduce search times in vivo. To model how changes to motifs and flanking sequences alter TF search behavior, we expanded our two-state model (in which a single TF is either bound or not bound to any location within the DNA; Fig. 2A) to a four-state continuous-time Markov Chain (CTMC) model in which a single TF may be (i) free (nonspecifically diffusing in the nucleoplasm), (ii) testing (near DNA or nonspecifically bound to DNA), (iii) bound to a motif, or (iv) bound to the flanks (Fig. 4D; see the materials and methods). The rate constant for transitioning between the free and testing states is given by $k_{\text{on,max}}$ (the theoretical upper bound for the on-rate if all nonspecific TF-DNA interactions result in specific binding); rate constants for transitioning from the motif- or flank-bound state to the testing state are given by $k_{\text{off,}\mu\text{,motif}}$ and $k_{\text{off,}\mu\text{,flank}}$; and the probabilities of transitioning to the motif or flanks depend on the likelihood of binding either sequence (f_{flank} or f_{motif}) and on the rate at which TFs transition from the testing state back to the free state ($k_{\text{off,M}}$). Together, this yields a simple expression for the transition probability from the testing state to either the flank or motif ($p_{\text{testing,x}} = f_x/(1 + f_{\text{flank}} + f_{\text{motif}})$; x ∈ {flank,motif}). Assuming that the time spent in the testing state is negligible, this four-state model can determine these microscopic rate constants from macroscopic measurements of affinities and apparent dissociation rates for sequences containing a consensus E-box,

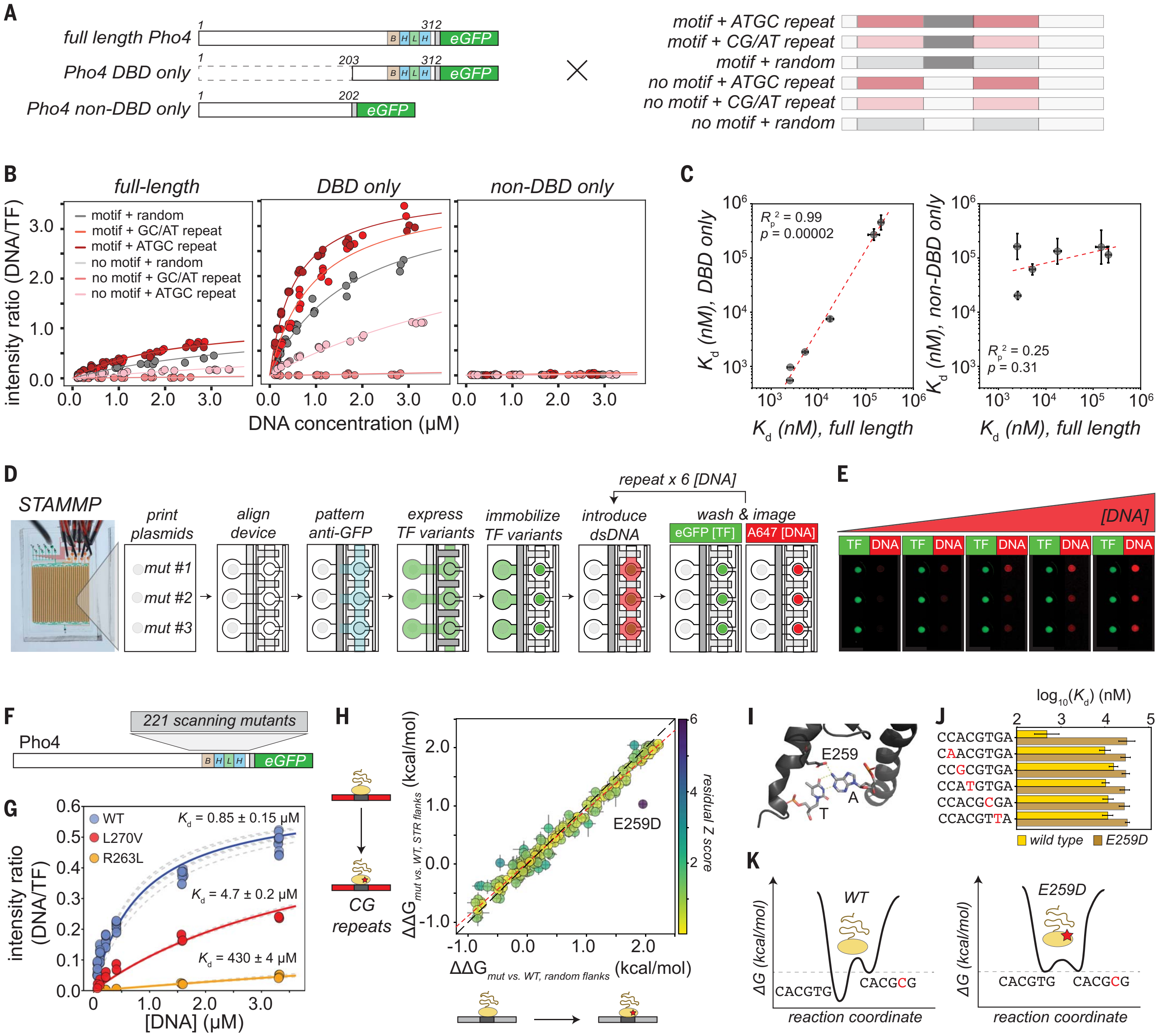


Fig. 3. Mutations within TF DBDs alter repeat sensitivity. (**A**) Schematic illustrating MITOMI experiment quantifying binding of full-length, DBD-only, and non-DBD–only Pho4 constructs to DNA sequences either containing or lacking motifs surrounded by either random sequence or favorable STRs. bHLH indicates the bHLH DBD within Pho4. (**B**) Measured concentration-dependent binding for full-length, DBD only, and non-DBD–only Pho4 constructs. Markers denote measured intensities from individual chambers; lines indicate Langmuir isotherm fits. (**C**) Scatter plots comparing measured K_ds for DBD-only Pho4 versus full-length Pho4 (left) and non-DBD–only Pho4 versus full-length Pho4 (right). Marker bars indicate mean across all chambers, error bars indicate SD, dashed line indicates linear regression, and *P* values indicate the significance of correlation. (**D**) Experimental pipeline for STAMMP illustrating steps for recombinant protein expression, surface immobilization, purification, and measurement of concentration-dependent binding behavior. (**E**) Example zoomed-in fluorescence images showing immobilized TFs and concentration-dependent DNA binding. (**F**) Schematic of C-terminally eGFP-tagged Pho4 and location of scanning mutants. (**G**) Example concentration-dependent binding measurements and Langmuir isotherm fits for WT Pho4 and two mutants (L270V and R263L) interacting with "Motif + random 1." (**H**) Effects of TF mutations on relative DNA-binding affinity for an extended E-box consensus flanked by CG repeats versus random sequence. Black dashed line indicates 1:1 relationship, red dashed line indicates linear regression, and color bar indicates *Z* score of residuals from linear regression. (**I**) Zoomed-in crystal structure showing contacts between the WT E259 and E-box consensus (PDB ID: 1a0a). (**J**) Affinities for Pho4 WT and E259D mutants interacting with consensus E-box and five single-nucleotide substitutions. (**K**) Reaction coordinate diagram of binding specificity landscapes for Pho4 WT and E259D.

a weak E-box, or a scrambled sequence surrounded by 13-bp flanks composed of either GT/AC or CG/AT dinucleotide repeats or random sequence (DNA Library 1; Fig. 4E and figs. S89 to S97; see the materials and methods). Consistent with recent work on *E. coli* LacI binding to various operator sequences ([79]), mean microscopic dissociation rates ($k_{off,\mu}$) for sequences with a consensus E-box or a weak E-box were similar, but affinities and microscopic association likelihoods differed by 12- or 16-fold (figs. S98 to S100). Systematically

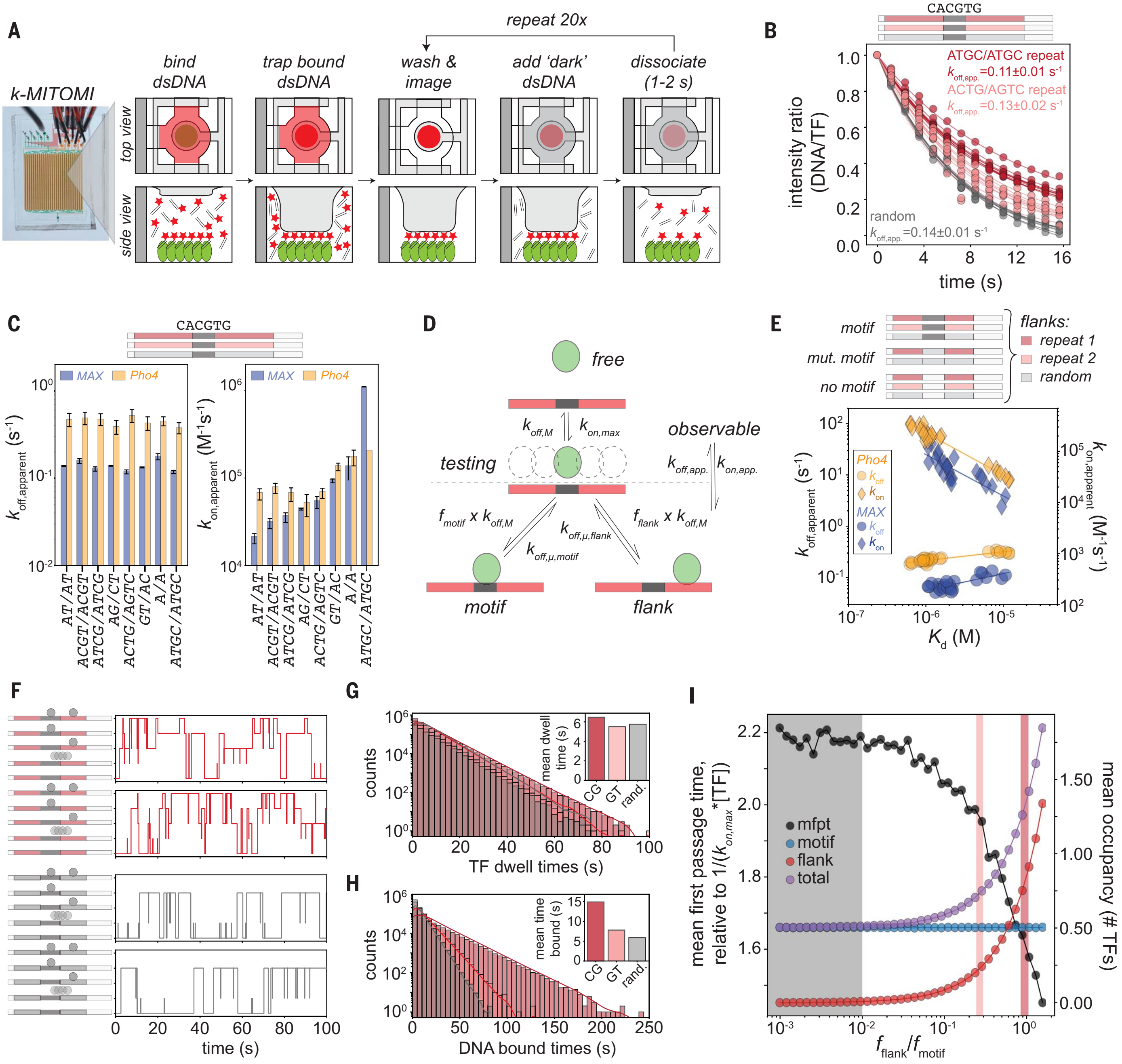


Fig. 4. Repetitive flanking sequences increase macroscopic association rates and reduce mean first passage time. (**A**) Experimental pipeline for *k*-MITOMI (see the materials and methods). (**B**) Example dissociation curves for MAX interacting with DNA Library 2 sequences showing per-chamber measurements (markers), per-chamber single-exponential fits (lines), and the average of returned fit parameters (annotation) for each sequence. (**C**) Measured $k_{off,apparent}$ (left) and calculated $k_{on,apparent}$ (right) values as a function of flanking sequence for Pho4 (yellow) and MAX (blue) interacting with DNA Library 2 sequences (all of which contain a core motif). (**D**) Proposed four-state model and associated microscopic rate constants for TF binding to sequences with a central core motif surrounded by different flanking sequences. (**E**) Average measured $k_{off,apparent}$ (circle markers, left axis) and calculated $k_{on,apparent}$ (diamond markers, right axis) values versus measured affinities (K_ds) for Pho4 (yellow) and MAX (blue) interacting with all sequences from DNA Library 1. (**F**) Sample TF trajectories from Gillespie simulations modeling 2600 TFs interacting with a single DNA sequence containing a consensus motif flanked by either repetitive (top, red) or random (bottom, gray) flanks. DNA can be unbound, associated with TFs in a "testing" state, bound by a TF at the motif, bound by a TF at the flanking sequence, or bound by TFs at the motif and flanking sequence simultaneously. (**G**) Log-linear distribution of TF dwell times across 1000 simulations for sequences with a consensus motif flanked by CG repeats, GT repeats, or random sequence. Inset shows mean dwell times by sequence. (**H**) Log-linear distribution of the lengths of time a DNA sequence is bound by at least one TF across 1000 simulations for sequences with a consensus motif flanked by GC repeats, GT repeats, or random sequence. Inset shows mean time occupied by sequence. (**I**) Mean first passage time {black markers, left axis; units relative to fastest possible search time, $1/[k_{on,max}$*(TF)]}, mean motif occupancy (blue markers, right axis), mean flank occupancy (red markers, right axis), and mean total DNA occupancy (purple markers, right axis) as a function of the likelihood of binding flanking sequence. Gray box indicates the range of affinities for random flanks; pink and red boxes correspond to f_{flank} values for GT and CG repeats, respectively.

quantifying how changes in microscopic rate constants affect macroscopic observables (K_d and $k_{\mathrm{off,apparent}}$) reveals combinations that can differentially affect affinity and kinetics [e.g., altering $k_{\mathrm{on,max}}$ does not change overall dissociation rates but can alter affinity when microscopic dissociation from the motif ($k_{\mathrm{off,\mu,motif}}$) is slow; figs. S101 and S102]. Because fitted microscopic rate parameters are often found at locations in phase space where concomitant variation in two parameters differentially tunes binding (figs. S101 and S102), STRs may maximize regulatory tunability (32, 80–83).

Using these microscopic rate parameters in Gillespie stochastic simulations to predict behavior for 2600 TFs [the estimated number of Pho4 copies in *S. cerevisiae* (84)] binding DNA within the yeast nucleus yielded individual TF trajectories that recapitulated the observed experimental trends (fig. S103) and showed that sequences with favorable flanking STRs were frequently occupied by multiple TFs (Fig. 4F and figs. S104 and S105). Although the DNA dwell time for any individual TF was largely independent of flanking sequence identity (Fig. 4G and fig. S106A), as expected with the absence of an observed macroscopic off-rate effect (Fig. 4, C and E, and fig. S87), DNA sequences with preferred flanking STRs were occupied by at least one TF for substantially more time (Fig. 4H and fig. S106B). Mean behavior across 100 simulations showed that as the relative affinity for flanking STRs increased, total DNA occupancy increased, creating a locally concentrated pool of TFs (Fig. 4I and fig. S106C). Although variations in the relative motif/flank affinity ratio did not affect the mean first passage time (MFPT) to the motif (the mean time for a TF to move from the free state to the motif state) or motif occupancy (as expected), changing this ratio altered the effective TF concentrations at which flanks were occupied (figs. S107 and S108). Even for this simple model that does not consider proximity between the motif and flanks, favorable STRs thus reduce MFPT to the entire DNA sequence (motif and flanking sequences) (Fig. 4I and figs. S109 and S110), consistent with a hypothesized role for STRs in regulating stress responses (3) and with previous work showing that favorable STRs can act as "antennae" to enhance TF target search (85).

STRs alter gene expression by tuning TF occupancies in vivo

Although STRs have repeatedly been associated with changes in gene expression in cells, and the length of STRs in the genome exceeds the length required for an in vitro effect (figs. S4 and S23), our results thus far did not elucidate whether STRs alter TF occupancies in vivo. Directly quantifying the impacts of STRs on TF binding in cells is technically challenging, because the lower-affinity binding expected for STRs is unlikely to yield distinct peaks within chromatin immunoprecipitation data. To sensitively quantify effects of STRs in vivo, we trained the BPNet (37) neural network (NN) on in vivo chromatin immunoprecipitation sequencing (ChIP-seq) data with 5-fold cross-validation to predict TF binding profiles from DNA sequence with nucleotide resolution and then applied Affinity Distillation (AD) (86) to predict log-transformed mean read counts [Δlog(counts)], which were previously shown to correlate with measured thermodynamic energies (ΔΔGs). If STRs alter gene expression in vivo by changing TF occupancies, then we would expect BPNet to learn that they affect TF binding and AD to predict sequence-dependent read count changes that mirror ΔΔGs measured in vitro.

After training on high-quality MAX ChIP-seq data (87, 88) (Fig. 5A), BPNet accurately predicted log-transformed read counts for held-out data (R^2 = 0.52), with binding profiles that reproduced those observed experimentally (Fig. 5A) (86). Returned contribution weight matrices (CWMs), which identify short subsequences most predictive of TF binding, revealed E-box–like motifs (CACGTG) that sometimes included a flanking preference for CG dinucleotides, consistent with in vitro preferences (Figs. 5A, 1G, and 2, G and H). Some CWMs also included an AP1-binding motif (TGACTCA), consistent with AP1 acting as a pioneer factor to increase chromatin accessibility for MAX (Fig. 5A) (89). AD-predicted 8-mer Z score distributions showed higher correlation with distributions calculated from uPBM data relative to either mononucleotide or dinucleotide models ($R_\mathrm{p}{}^2$ = 0.42, 0.21, and 0.22, respectively), likely because of an enhanced ability to accurately predict low-affinity interactions (figs. S111 to S114). AD-predicted log-transformed read counts for DNA Library 1 sequences also strongly correlated with measured ΔΔGs (R^2 = 0.78) and partition function–predicted binding energies [Fig. 5B and figs. S115 to S118; see additional data at (55)]. AD consistently predicted tighter binding to consensus motifs flanked by preferred STRs (Fig. 5B), and importance scores from DeepSHAP (90, 91), which identify base pair contributions to the observed model output, confirmed that enhanced binding was caused by the flanking STRs in these synthetic sequences (Fig. 5, C and D, and figs. S119 to S121). Together, these analyses suggest that observed in vivo effects of polymorphic STRs on gene expression can be explained at least in part by differential TF binding.

STR impacts extend over tens of nucleotides and mismatches reduce effects

To determine the distance over which STRs affect binding, we quantified MAX binding affinities for DNA containing an E-box motif surrounded by increasing lengths (15, 30, 45, or 60 bp) of either disfavored (AG/CT) or favored (GT/AC) repeats using MITOMI (table S6). In parallel, we used AD to predict MAX occupancies and binding profiles for the same sequences (Fig. 5, E to G, and fig. S122). For disfavored AG/CT repeats, both MITOMI and AD revealed that increasing STR lengths monotonically reduced binding, with effects saturating after ~40 bp (Fig. 5G; R^2 = 0.80 between predictions and measured ΔΔGs). Returned DeepSHAP interpretations and cumulative importance scores confirmed a negative contribution from flanking STRs (Fig. 5, F and H). Favored GT/AC repeats showed more complex behavior, with short repeats (15 to 30 bp) increasing binding and longer repeats having only minor effects, but predictions were again consistent with experimental observations (fig. S122; R^2 = 0.93).

Nearly 80% of repeated units within the median human STR match the consensus repeat exactly, with the remaining 20% containing an indel or mismatched base(s) (see the materials and methods). To investigate how imperfections within STRs alter binding, we applied MITOMI and AD to measure and predict MAX binding to seven increasingly scrambled (GT/AC) repeat sequences (Fig. 5I). Even though the relationship between measured affinities and repeat imperfection (as quantified by Shannon entropy) was nonmonotonic, AD accurately predicted energetic measurements (R^2 = 0.84), suggesting that the algorithm had learned that the increased multiplicity of even weakly preferred STRs enhances binding and that energetic impacts depend not only on nucleotide composition but also on repeat imperfection (Fig. 5, J and K).

TF binding to STRs is widespread across structural families and organisms

To determine whether STR binding is specific to Pho4 and MAX or more widespread, we analyzed PBM data for 1291 TFs from 114 species, including *S. cerevisiae*, *Arabidopsis thaliana*, *D. melanogaster*, *Caenorhabditis elegans*, *M. musculus*, and *H. sapiens* (61, 64, 92) (table S15). For each experiment for each TF, we iterated through all 65,536 (4^8) 8-mers, computed median intensities for all probes containing each 8-mer, and calculated Z scores relative to this distribution for all 39 nonredundant homopolymeric, dinucleotide repeat, and tetranucleotide 8-mer STRs (Fig. 6A, figs. S123 and S124; see the materials and methods). TF preference for STRs was ubiquitous, with 90% (1158/1291) of all TFs binding at least one STR with $P < 1.3 \times 10^{-3}$ (the Bonferroni-corrected threshold for significance; figs. S125 and S126 and table S15), and STR preferences varied widely across TF families (figs. S127 to S143). Some families (e.g., nuclear hormone receptors, T-box, and bZIP) show little preference for any STRs, whereas others (e.g., AT

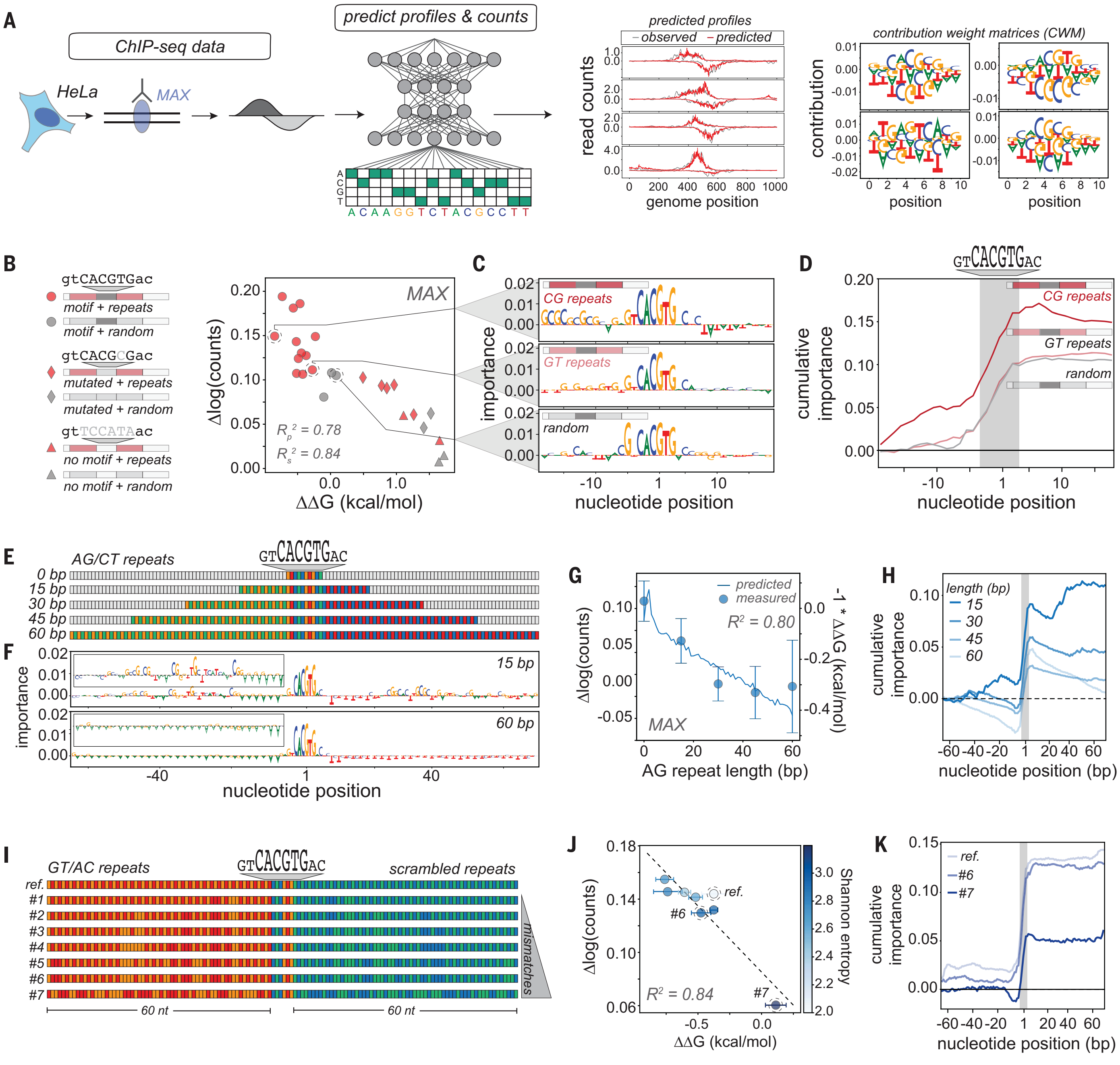


Fig. 5. NN models trained on in vivo datasets recapitulate repeat effects observed in vitro and return predictions similar to statistical mechanics models. (**A**) Experimental pipeline: AD NN models trained on MAX ChIP-seq data predict base pair–resolution binding profiles and return hypothetical CWMs representing binding preferences. Positive and negative numbers represent nucleotides that favor and disfavor binding, respectively. (**B**) AD-predicted binding [Δlog(counts)] versus MITOMI-measured ΔΔGs for 26 DNA sequences containing an intact motif, a mutated motif, or scrambled sequence surrounded by either repetitive (red markers) or random (gray markers) flanking sequence. (**C**) DeepSHAP interpretations for a motif surrounded by a favored repeat (CG, top), a disfavored repeat (GT, middle), or random sequence (bottom). The sum of importance scores across a sequence are equal to the count prediction output of the NN. (**D**) Cumulative importance scores as a function of position for a favored repeat (CG, dark red), a disfavored repeat (GT, light red), or random sequence (gray). Gray box indicates motif location. (**E**) Schematic of sequences with E-box and 15, 30, 45, or 60 bp of disfavored AG/CT repeats. (**F**) DeepSHAP interpretations for 15- and 60-bp sequences from (E). (**G**) AD-predicted change in log(counts) (blue line, left axis) and −1*MITOMI-measured ΔΔGs (blue markers, right axis) as a function of repeat length (relative to a sequence with a motif and random flanks). Markers and error bars show median and SD across replicates, respectively. (**H**) Cumulative importance scores as a function of position for sequences with E-box and 15, 30, 45, or 60 bp of AG/CT repeats. Gray box indicates motif position. (**I**) Schematic of sequences with E-box and increasingly scrambled GT/AC repeats. (**J**) AD-predicted change in log(counts) (blue line, left axis) and −1*MITOMI-measured ΔΔGs (blue markers, right axis) for sequences shown in (I) calculated relative to reference sequence. Color indicates Shannon entropy. Markers and error bars show median and SD across replicates, respectively. (**K**) Cumulative importance scores as a function of position for reference sequence and sequences 6 and 7. Gray box indicates motif position.

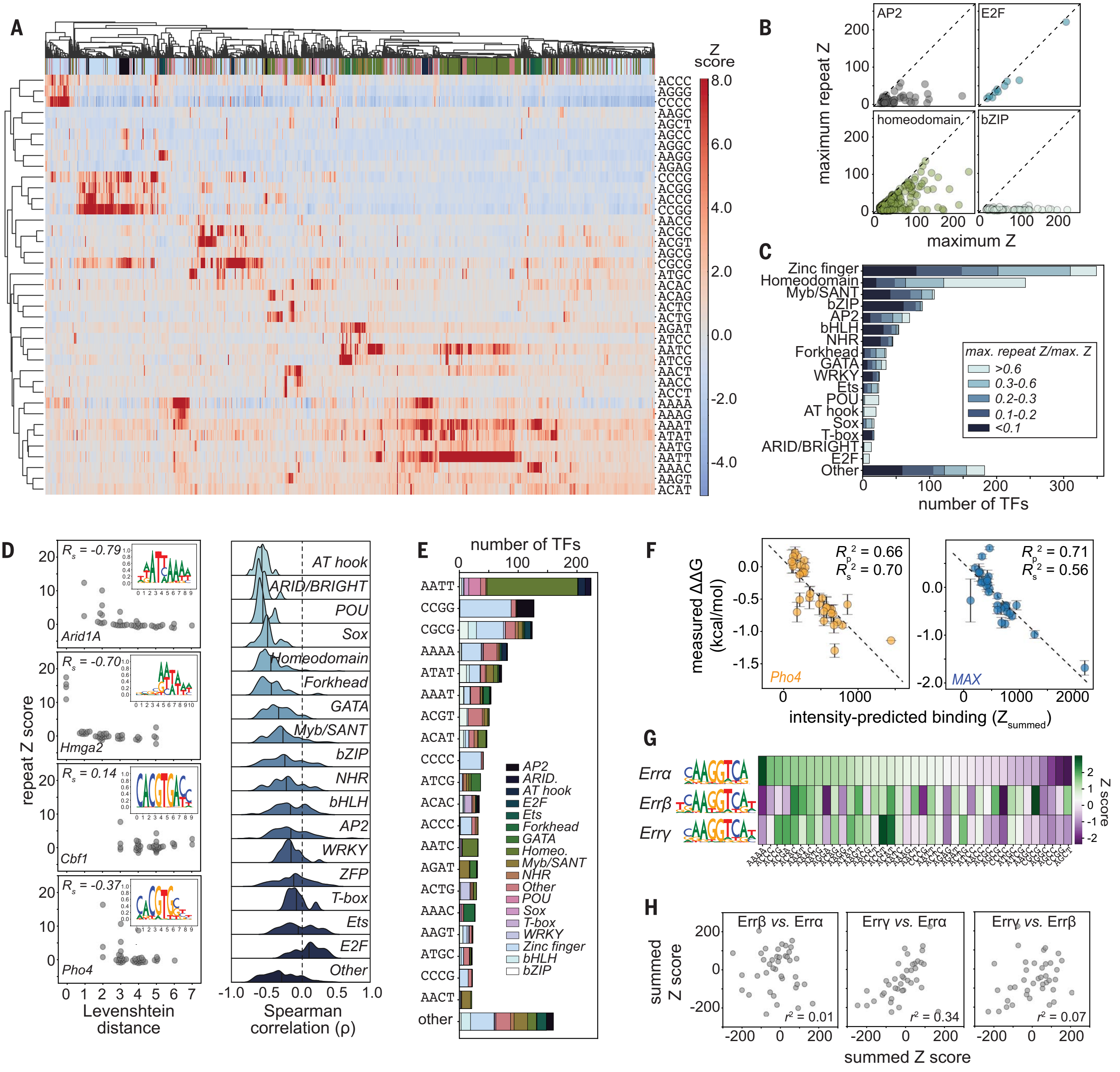


Fig. 6. Most TFs show statistically significant binding to repetitive sequences. (**A**) Heatmap showing calculated 8-mer intensity Z scores for 1291 TFs (columns) interacting with 39 nonredundant STR types (rows; i.e., reverse complements are considered a single sequence). (**B**) Maximum repeat Z score versus maximum overall Z score for TFs from four different structural families: AP2, E2F, homeodomain, and bZIP. (**C**) Distributions of ratios of maximum repeat Z scores relative to maximum overall Z scores across TF families. (**D**) Left: Repeat Z score as a function of Levenshtein distance from preferred consensus sequence for Arid1a, Hmga2, Cbf1, and Pho4. Insets show PWM representations of preferred consensus sequences downloaded from CIS-BP (64). Right: Distributions of Spearman correlation coefficients between repeat Z score and Levenshtein distance from consensus across 17 different TF structural families. (**E**) Bar plot showing the number of TFs that prefer a particular tetranucleotide repeat shaded by TF family. (**F**) Scatter plots, linear regressions, and correlation coefficients for measured ΔΔGs versus summed Z scores (intensity-predicted binding) across all measured repeats for Pho4 (left) and MAX (right). (**G**) PSAM (left) and heat map showing 8-mer Z scores for three NHR paralogs from *M. musculus* (Errα, Errβ, and Errγ). (**H**) Pairwise comparisons of predicted binding (calculated by summing Z scores) for consensus motifs surrounded by 50 bp on either side of the tetranucleotide repeats to Errα, Errβ, and Errγ.

hook, E2F, and ARID/BRIGHT families) prefer STRs simply because they resemble the known consensus (Fig. 6, B and C). Members of multiple families (e.g., AP2, Forkhead, GATA, homeodomain, Myb/SANT, zinc fingers, and bHLH) weakly preferred particular STRs (Fig. 6, B and C) that often have little sequence similarity to the known motif (as quantified by Levenshtein distance; Figs. 6D and fig. S144). Across all TFs, AATT and CCGG repeats were the most preferred, largely because these STRs resemble known motifs for TFs in two most abundant structural families [homeodomain and zinc finger TFs, respectively (93)] (Fig. 6E). C homopolymers were the most disfavored (fig. S145).

Differential STR preferences could allow closely related paralogs to target distinct genes

Many closely related paralogs with conserved DBDs and nearly identical consensus motif preferences bind and regulate distinct gene targets in vivo (94–96). This differential binding has been attributed to either subtle differences in motif (65) or flanking nucleotide (57, 67, 97–99) preferences or direct binding by poorly conserved regions outside of the DBD (72). As an alternate hypothesis, differential STR preferences could influence paralog-specific localization. Global comparisons of preferred STRs and preferred motifs across paralogs within a species (quantified through cosine similarity; see the materials and methods) revealed many TF pairs with highly similar motifs but divergent STR preferences (figs. S146 to S151), particularly for bHLH and nuclear hormone receptor (NHR) TF paralogs in *A. thaliana* and *M. musculus* (figs. S152 to S154).

Uncalibrated summed 8-mer Z scores for Pho4 and MAX binding to DNA Library 2 sequences correlated well with measured ΔΔGs (R^2 = 0.66 and 0.71 for Pho4 and MAX, only slightly worse than for calibrated partition function–based predictions) (Fig. 6F), suggesting that existing PBM measurements can be used to estimate binding to arbitrary sequences even without quantitative affinity measurements. Predicted binding of the Errα, Errβ, and Errγ NHR TFs from *M. musculus* (which have nearly identical motifs but distinct STR preferences) to sequences containing the consensus surrounded by 50 bp (on either side) of random sequence or STRs were poorly correlated (R^2 = 0.01, 0.34, and 0.07, respectively; Fig. 6, G and H), consistent with the hypothesis that sensitivity to STRs could differentially localize paralogs.

STRs are associated with active enhancers and high mutation rates

STRs can enhance or decrease TF binding energies; however, the lower bound of affinity imposed by nonspecific, electrostatic-mediated interactions skews STR effects to predominantly enhance binding (fig. S124). Consistent with a primarily activating role, STRs are most enriched within the most active enhancers [$R_S{}^2$ = 0.67, as measured by CAGE-seq, p300 ChIP, GRO-seq, or similar enhancer activity assays (100); control datasets shuffling enhancer sequences and measured activity show no significant correlation ($R_S{}^2$ = 0.16)] (Fig. 7A). STRs are also preferentially enriched in enhancers that are broadly active across 278 human cell types ($R_S{}^2$ = 0.85); shuffled negative control datasets show no enrichment ($R_S{}^2$ = 0.02) (Fig. 7B). Across various eukaryotic genomes, mutations in STRs occur several orders of magnitude more frequently than short insertions and deletions (indels, 1 to 3 bp) and base substitutions (Fig. 7C), suggesting that STRs can provide an easily evolvable mechanism to tune transcription (3, 12, 101).

Discussion

The role of STRs in transcriptional regulation has been thoroughly documented, yet the mechanism by which they alter gene expression is poorly understood. Here, we present a model in which STRs directly bind TFs, thus establishing STRs as a class of regulatory elements. Our model is consistent with prior work suggesting that STRs tune gene expression by modulating nucleosome occupancy (3), because TF binding, especially that of pioneer factors, is the primary determinant of chromatin accessibility (102–105). However, this model allows for more sophisticated regulation: Rather than uniformly altering chromatin accessibility, STRs can differentially affect binding for even closely related TFs, serving as rheostats to precisely tune TF binding at a specific locus (81, 83, 106–108). Moreover, with relatively few types of STRs relative to the number of different TFs, STRs in the absence of known motifs can recruit a diverse set of TFs, thereby functioning as general regulatory elements, consistent with observations that STR-enriched enhancers are broadly active across cell types (27) (Fig. 7). Finally, STRs need not surround a TF consensus motif to have a regulatory effect; rather, they may sequester TFs for precise temporal control of transcription, as is hypothesized for pericentromeric satellites regulating the timing of chromosomal replication (109).

In contrast to the canonical model that long residence times confer specificity and function whereas TF search is nonspecific and diffusion limited (110), we found that favorable STRs surrounding target motifs alter affinities primarily by increasing apparent macroscopic TF association rates. These results contradict prior measurements suggesting that DNA sequence variation primarily affects dissociation rates; however, prior experiments did not include unlabeled competitor DNA and therefore likely observed a convolved process of dissociation and rebinding (110, 111). Thus, we join other recent work in challenging the canonical view that protein–nucleic acid binding affinities are primarily determined by dissociation rates (79). Our measurements can be explained by a simple four-state model showing that STRs enhance affinities by increasing the rate of DNA association. This is consistent with prior work suggesting that degenerate recognition sites may serve as “DNA antennae” to attract TFs to a particular regulatory site (85, 112–115). This four-state model likely underestimates the true impacts of STRs on target search because it does not explicitly consider whether TFs can move from flanking STRs to a central motif through one-dimensional sliding, hopping, and intersegmental transfer (116–119), rather than dissociating, diffusing, and rebinding. Future experiments will be required to deconvolve the kinetic contributions of nonspecific, electrostatic-mediated binding from other “testing” states for different TF structural classes, to quantify the effects of facilitated dissociation on the observed macroscopic and inferred microscopic parameters (111, 120, 121), and to develop more complex models that consider the contribution of each microstate to macroscopic kinetic parameters.

Because eukaryotic TFs recruit transcriptional coactivators through “fuzzy,” multivalent (122–124), and allovalent (71) interactions, the finding that STRs enhance the local concentration and reduce mean first passage time near genomic target sites raises the intriguing possibility that dense clusters of loosely bound TFs could enhance the recruitment of coactivator proteins to ensure fast transcriptional response kinetics. This hypothesis is supported by the observation that STRs in budding yeast are enriched near binding sites of stress response TFs (3), for which a rapid transcriptional response may be especially advantageous. The smaller size and operon structure within bacterial and archaeal genomes suggests a reduced need to speed TF search. Consistent with this, bacterial TFs tend to bind long target motifs with high affinities (108, 125, 126), and STRs comprise a smaller percentage of bacterial and archaeal genomes (table S16 and figs. S155 and S156).

This case study of STRs further underscores the limitations of motif-based models in predicting TF occupancy from sequence (37, 86, 127–130), because STRs composed of overlapping instances of even low-affinity sites bearing little resemblance to the known motif can substantially alter binding. Binding of the same TF to dissimilar motifs has previously been reported and attributed to alternate binding modes driven by either entropic or enthalpic effects (131–133). Although previous reports have identified repeated instances of motif and motif-like sequences that bind TFs and thereby alter gene expression (20, 134, 135), these observations are well explained by simple position weight matrix models (136–138) that do not predict the enhanced binding to STRs observed here. Here, we show that statistical mechanical models that explicitly account for low-affinity binding substantially improve quantitative binding predictions for arbitrary DNA sequences relative to motif-based approaches. While this effect is particularly apparent for STRs, we also expect nonrepetitive sequence contexts containing many low-affinity binding sites to show similar effects. In future work, small sets of absolute affinity measurements across many TFs could be combined with statistical mechanical and machine learning models to enable quantitative predictions

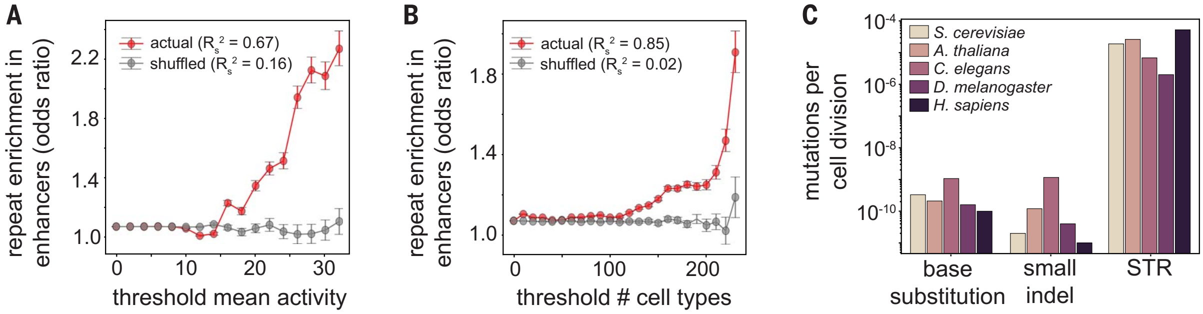


Fig. 7. STRs are associated with active enhancers and have high mutation rates. (**A**) Enrichment of STRs in enhancers versus shuffled negative controls as a function of mean enhancer activity. Error bars are 95% confidence intervals. (**B**) Enrichment of STRs in enhancers versus shuffled negative controls as a function of the number of cell types within which an enhancer is active. Error bars are 95% confidence intervals. (**C**) Calculated rate of mutation (per cell division) for base substitutions, small indels, and STRs in five different model organisms.

of how changes in nuclear TF concentration alter cooperation and competition between TFs to drive specific transcriptional programs.

Because our statistical mechanics framework is agnostic to the identity of binding partners and considers only a distribution of binding energies, we anticipate that the same physical considerations by which DNA-binding proteins recognize STRs may also apply to RNA-binding proteins. Evidence in the literature already points to a role for intronic STRs in regulating splicing (139–150) or promoting the formation of RNP compartments (151–153). These observations raise the intriguing possibility that STR-enriched enhancers could serve a dual function of binding TFs to regulate transcription and subsequently recruiting RNA-binding proteins once transcribed into enhancer RNAs.

STRs are highly evolvable (101, 154), requiring only mispairing during replication, repair, or recombination to expand or contract (155–157), and may therefore serve as the raw material for evolving new *cis*-regulatory elements (101, 158) and fine-tuning existing regulatory modules for sensitive transcriptional programs, such as those in development (159). This work may motivate future efforts to assess the evolution of regulatory networks across species by considering not only conservation of nucleotides within motifs, but also the types and lengths of STRs surrounding them. The evolution of regulatory STRs is likely complemented by the coevolution of TF binding preferences, consistent with a model in which DBDs exist as a conformational ensemble of partially folded states in which single-residue substitutions alter the distribution of states within the ensemble and therefore tune the specificity or promiscuity of binding (50, 160–163). The observation that STR polymorphisms disrupt gene expression by directly altering TF binding may provide new clinical insights and therapeutic directions for a variety of STR-associated diseases, from autism (29, 30) to microsatellite instability–associated cancers (164, 165) and others yet to be discovered.

Methods summary

Microfluidic device fabrication and operation

Microfluidic devices were fabricated and aligned to printed oligonucleotide or plasmid DNA arrays as described previously (48, 50). Microfluidic devices were controlled by a custom pneumatic manifold (166) and imaged with a fully automated microscope and custom software (50, 160).

MITOMI and k-MITOMI experiments

Single-stranded DNA oligonucleotide libraries were synthesized by Integrated DNA Technologies (IDT) and fluorescently labeled and duplexed with a primer extension step. eGFP-tagged TFs were expressed off-chip with wheat germ extract or PURExpress (New England Biolabs) and purified with anti-eGFP antibodies on the device. Printed fluorescent DNA was solubilized in TBS or wheat germ extract and allowed to bind to immobilized TFs for 90 min before washing out unbound species and imaging. Binding was quantified as the ratio of DNA fluorescence to TF fluorescence, and the resulting data for multiple concentrations of DNA were fit to a Langmuir isotherm to extract K_d and ΔΔG values. For kinetic measurements, excess unlabeled ("dark") dsDNA was iteratively introduced in solution, and button valves were opened to allow dissociation. Macroscopic dissociation rates ($k_{\text{off,apparent}}$) were fit to the ratio of DNA fluorescence to TF fluorescence over several time points to an exponential decay. Apparent macroscopic association rates ($k_{\text{on,apparent}}$) were inferred by $k_{\text{on,apparent}} = k_{\text{off,apparent}}/K_d$, assuming a two-state macroscopic binding model.

Partition function models of binding

Partition function–based models of binding are based on 8-mer intensities derived from previously published uPBM data. uPBM data and associated *Z* scores for all possible 8-mers were downloaded from CIS-BP (64) and filtered for data quality. We predicted relative binding energies for an arbitrary sequence to a given TF by splitting the sequence into overlapping 8-mers and computing the following: $\Delta\Delta G = c\beta\log\sum_j e^{\beta\log I_j} - \Delta G_{ref}$, where β = 1/($k_B T$), I_j is the PBM intensity for an 8-mer *j*, and *c* is some calibration constant determined by a linear fit between PBM and MITOMI data.

STAMMP experiments

Single-stranded DNA oligonucleotides were synthesized by IDT and fluorescently labeled and duplexed with a primer extension step. eGFP-tagged TFs were expressed and purified on-chip with PURExpress, and increasing concentrations of fluorescently labeled dsDNA were flowed over the chip and allowed to bind for 50 min before washing and imaging. Binding was quantified as a ratio of DNA fluorescence to TF fluorescence, and the resulting data for multiple concentrations of DNA were fit to a Langmuir isotherm to extract K_d and ΔΔ*G* values.

CTMC and Gillespie models

Microscopic kinetic parameters were fit to a four-state CTMC kinetic model in which a TF can be free, nonspecifically bound and testing, bound to the motif, or bound to the flanks from mean $k_{\text{off,apparent}}$ and K_d measurements with a custom MATLAB script. Gillespie simulations were performed using custom Python scripts with microscopic parameters fit from the Pho4 CTMC model and 10,000 iterations per parameter set with 2600 TFs and 100,000 s per simulation.

Affinity distillation

ChIP-seq data for MAX were downloaded from the ENCODE portal (87, 88) with accession numbers ENCSR000EZM (control) and ENCSR000EZF (experiment). NN architecture was adapted from BPNet (37) and trained on IDR peaks, with regions from chromosomes 8 and 9 used as the test set and regions from chromosomes 16, 17, and 18 used as the tuning set for hyperparameter tuning. All NN models were implemented and trained in Keras (v.2.2.4; TensorFlow backend v.1.14) (*167*, *168*) using the Adam optimizer (*169*). AD scores [Δlog (counts)] were calculated by inserting a given sequence at the center of 100 different background sequences and computing the mean of the differences between the log(count) predictions for query sequence and background sequence alone, as described in (*86*).

Bioinformatic analyses

STRs in the human genome were identified using Tandem Repeats Finder (*170*). Genome annotations used to calculate enrichment of STRs in enhancers were downloaded from the Enhancer Atlas (*100*), FANTOM 5 (*171*), and HACER (*172*) databases. Mutation rates per cell division were cited or calculated as previously described (*154*, *173*–*175*).

REFERENCES AND NOTES

1. S. Nurk *et al*., The complete sequence of a human genome. *Science* **376**, 44–53 (2022). doi: 10.1126/science.abj6987; pmid: 35357919
2. E. S. Lander *et al*., Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). doi: 10.1038/35057062; pmid: 11237011
3. M. D. Vinces, M. Legendre, M. Caldara, M. Hagihara, K. J. Verstrepen, Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–1216 (2009). doi: 10.1126/science.1170097; pmid: 19478187
4. S. Sawaya *et al*., Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLOS ONE* **8**, e54710 (2013). doi: 10.1371/journal.pone.0054710; pmid: 23405090
5. M. Gymrek *et al*., Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016). doi: 10.1038/ng.3461; pmid: 26642241
6. A. Contente, A. Dittmer, M. C. Koch, J. Roth, M. Dobbelstein, A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30**, 315–320 (2002). doi: 10.1038/ng836; pmid: 11919562
7. H. Hamada, M. Seidman, B. H. Howard, C. M. Gorman, Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Mol. Cell. Biol.* **4**, 2622–2630 (1984). pmid: 6098815
8. F. Gebhardt, K. S. Zänker, B. Brandt, Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**, 13176–13180 (1999). doi: 10.1074/jbc.274.19.13176; pmid: 10224073
9. S. Shimajiri *et al*., Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett.* **455**, 70–74 (1999). doi: 10.1016/S0014-5793(99)00863-7; pmid: 10428474
10. K. M. Warpeha *et al*., Genotyping and functional analysis of a polymorphic (CCTTT)(n) repeat of NOS2A in diabetic retinopathy. *FASEB J.* **13**, 1825–1832 (1999). doi: 10.1096/fasebj.13.13.1825; pmid: 10506586
11. R. I. Richards, K. Holman, S. Yu, G. R. Sutherland, Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum. Mol. Genet.* **2**, 1429–1435 (1993). doi: 10.1093/hmg/2.9.1429; pmid: 8242066
12. A. Sulovari *et al*., Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23243–23253 (2019). doi: 10.1073/pnas.1912175116; pmid: 31659027
13. A. J. Hannan, Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018). doi: 10.1038/nrg.2017.115; pmid: 29398703
14. A. C. Johnson, Y. Jinno, G. T. Merlino, Modulation of epidermal growth factor receptor proto-oncogene transcription by a promoter site sensitive to S1 nuclease. *Mol. Cell. Biol.* **8**, 4174–4184 (1988). pmid: 2847030
15. B. Wang, J. Ren, L. L. P. J. Ooi, S. S. Chong, C. G. L. Lee, Dinucleotide repeats negatively modulate the promoter activity of Cyr61 and is unstable in hepatocellular carcinoma patients. *Oncogene* **24**, 3999–4008 (2005). doi: 10.1038/sj.onc.1208550; pmid: 15782120
16. A. Heidari *et al*., Core promoter STRs: Novel mechanism for inter-individual variation in gene expression in humans. *Gene* **492**, 195–198 (2012). doi: 10.1016/j.gene.2011.10.028; pmid: 22037607
17. R. Meloni, V. Albanèse, P. Ravassard, F. Treilhou, J. Mallet, A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Hum. Mol. Genet.* **7**, 423–428 (1998). doi: 10.1093/hmg/7.3.423; pmid: 9466999
18. Y.-H. Chen *et al*., Microsatellite polymorphism in promoter of heme oxygenase-1 gene is associated with susceptibility to coronary artery disease in type 2 diabetic patients. *Hum. Genet.* **111**, 1–8 (2002). doi: 10.1007/s00439-002-0769-4; pmid: 12136229
19. J. Margoliash, S. Fuchs, Y. Li, A. Massarat, A. Goren, M. Gymrek, Polymorphic short tandem repeats make widespread contributions to blood and serum traits. bioRxiv 502370 [Preprint] (2022); .doi: 10.1101/2022.08.01.502370
20. K. Gangwal *et al*., Microsatellites as EWS/FLI response elements in Ewing's sarcoma. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10149–10154 (2008). doi: 10.1073/pnas.0801073105; pmid: 18626011
21. H. Geng *et al*., Interaction between CA repeat microsatellites and HIF1α regulated the transcriptional activity of porcine IGF1 promoter. *J. Appl. Genet.* **61**, 105–112 (2020). doi: 10.1007/s13353-019-00529-4; pmid: 31673965
22. R. Srinivasan *et al*., Zscan4 binds nucleosomal microsatellite DNA and protects mouse two-cell embryos from DNA damage. *Sci. Adv.* **6**, eaaz9115 (2020). doi: 10.1126/sciadv.aaz9115; pmid: 32219172
23. M. Mellul *et al*., Repetitive DNA symmetry elements negatively regulate gene expression in embryonic stem cells. *Biophys. J.* **121**, 3126–3135 (2022). doi: 10.1016/j.bpj.2022.07.011; pmid: 35810331
24. Q. Lu, L. L. Wallrath, H. Granok, S. C. Elgin, (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. *Mol. Cell. Biol.* **13**, 2802–2814 (1993). pmid: 8474442
25. A. Orian *et al*., Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* **17**, 1101–1114 (2003). doi: 10.1101/gad.1066903; pmid: 12695332
26. J. T. Streelman, T. D. Kocher, Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiol. Genomics* **9**, 1–4 (2002). doi: 10.1152/physiolgenomics.00105.2001; pmid: 11948285
27. J. O. Yáñez-Cuna *et al*., Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24**, 1147–1156 (2014). doi: 10.1101/gr.169243.113; pmid: 24714811
28. A. J. Hannan, Tandem repeat polymorphisms: Modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.* **26**, 59–65 (2010). doi: 10.1016/j.tig.2009.11.008; pmid: 20036436
29. I. Mitra *et al*., Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021). doi: 10.1038/s41586-020-03078-7; pmid: 33442040
30. B. Trost *et al*., Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020). doi: 10.1038/s41586-020-2579-z; pmid: 32717741
31. S. F. Fotsing *et al*., The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019). doi: 10.1038/s41588-019-0521-9; pmid: 31676866
32. T. Raveh-Sadka *et al*., Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **44**, 743–750 (2012). doi: 10.1038/ng.2305; pmid: 22634752
33. V. Iyer, K. Struhl, Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995). doi: 10.1002/j.1460-2075.1995.tb07255.x; pmid: 7781610
34. C. Fiore, B. A. Cohen, Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res.* **26**, 778–786 (2016). doi: 10.1101/gr.200733.115; pmid: 27197208
35. F. Liu, J. W. Posakony, Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules. *PLOS Genet.* **8**, e1002796 (2012). doi: 10.1371/journal.pgen.1002796; pmid: 22792075
36. J. Erceg *et al*., Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLOS Genet.* **10**, e1004060 (2014). doi: 10.1371/journal.pgen.1004060; pmid: 24391522
37. Ž. Avsec *et al*., Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021). doi: 10.1038/s41588-021-00782-6; pmid: 33603233
38. L. M. Liberman, A. Stathopoulos, Design flexibility in cis-regulatory control of gene expression: Synthetic and comparative evidence. *Dev. Biol.* **327**, 578–589 (2009). doi: 10.1016/j.ydbio.2008.12.020; pmid: 19135437
39. R. W. Lusk, M. B. Eisen, Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLOS Genet.* **6**, e1000829 (2010). doi: 10.1371/journal.pgen.1000829; pmid: 20107516
40. I. Sela, D. B. Lukatsky, DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.* **101**, 160–166 (2011). doi: 10.1016/j.bpj.2011.04.037; pmid: 21723826
41. A. Afek, D. B. Lukatsky, Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophys. J.* **105**, 1653–1660 (2013). doi: 10.1016/j.bpj.2013.08.033; pmid: 24094406
42. A. Afek, J. L. Schipper, J. Horton, R. Gordân, D. B. Lukatsky, Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17140–17145 (2014). doi: 10.1073/pnas.1410569111; pmid: 25313048
43. A. R. Iglesias, E. Kindlund, M. Tammi, C. Wadelius, Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. *Gene* **341**, 149–165 (2004). doi: 10.1016/j.gene.2004.06.035; pmid: 15474298
44. A. Afek, H. Cohen, S. Barber-Zucker, R. Gordân, D. B. Lukatsky, Nonconsensus protein binding to repetitive DNA sequence elements significantly affects eukaryotic genomes. *PLOS Comput. Biol.* **11**, e1004429 (2015). doi: 10.1371/journal.pcbi.1004429; pmid: 26285121
45. A. Afek, I. Sela, N. Musa-Lempel, D. B. Lukatsky, Nonspecific transcription-factor-DNA binding influences nucleosome occupancy in yeast. *Biophys. J.* **101**, 2465–2475 (2011). doi: 10.1016/j.bpj.2011.10.012; pmid: 22098745
46. M. Goldshtein *et al*., Transcription factor binding in embryonic stem cells is constrained by DNA sequence repeat symmetry. *Biophys. J.* **118**, 2015–2026 (2020). doi: 10.1016/j.bpj.2020.02.009; pmid: 32101712
47. S. J. Maerkl, S. R. Quake, A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007). doi: 10.1126/science.1131007; pmid: 17218526
48. P. M. Fordyce *et al*., De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol.* **28**, 970–975 (2010). doi: 10.1038/nbt.1675; pmid: 20802496
49. M. Geertz, D. Shore, S. J. Maerkl, Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16540–16545 (2012). doi: 10.1073/pnas.1206011109; pmid: 23012409
50. A. K. Aditham, C. J. Markin, D. A. Mokhtari, N. DelRosso, P. M. Fordyce, High-throughput affinity measurements of transcription factor and DNA mutations reveal affinity and specificity determinants. *Cell Syst.* **12**, 112–127.e11 (2021). doi: 10.1016/j.cels.2020.11.012; pmid: 33340452
51. E. M. O'Neill, A. Kaffman, E. R. Jolly, E. K. O'Shea, Regulation of PHO4 nuclear localization by the PHO80-PHO85 cyclin-CDK complex. *Science* **271**, 209–212 (1996). doi: 10.1126/science.271.5246.209; pmid: 8539622
52. K. Vogel, W. Hörz, A. Hinnen, The two positively acting regulatory proteins PHO2 and PHO4 physically interact with PHO5 upstream activation regions. *Mol. Cell. Biol.* **9**, 2050–2057 (1989). pmid: 2664469
53. C. Bouchard, P. Staller, M. Eilers, Control of cell proliferation by Myc. *Trends Cell Biol.* **8**, 202–206 (1998). doi: 10.1016/S0962-8924(98)01251-3; pmid: 9695840
54. A. Cascón, M. Robledo, MAX and MYC: A heritable breakup. *Cancer Res.* **72**, 3119–3124 (2012). doi: 10.1158/0008-5472.CAN-11-3891; pmid: 22706201
55. C. A. Horton *et al*., Data for: Short tandem repeats bind transcription factors to tune eukaryotic gene expression, Zenodo (2023); https://zenodo.org/record/8161431.

56. R. Rohs *et al*., The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248–1253 (2009). doi: 10.1038/nature08473; pmid: 19865164
57. R. Gordân *et al*., Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* **3**, 1093–1104 (2013). doi: 10.1016/j.celrep.2013.03.014; pmid: 23562153
58. M. A. H. Samee, B. G. Bruneau, K. S. Pollard, A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs. *Cell Syst.* **8**, 27–42.e6 (2019). doi: 10.1016/j.cels.2018.12.001; pmid: 30660610
59. S. Pal, J. Hoinka, T. M. Przytycka, Co-SELECT reveals sequence non-specific contribution of DNA shape to transcription factor binding in vitro. *Nucleic Acids Res.* **47**, 6632–6641 (2019). doi: 10.1093/nar/gkz540; pmid: 31226207
60. T. Zhou *et al*., Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4654–4659 (2015). doi: 10.1073/pnas.1422023112; pmid: 25775564
61. M. F. Berger *et al*., Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006). doi: 10.1038/nbt1246; pmid: 16998473
62. G. Badis *et al*., Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009). doi: 10.1126/science.1162327; pmid: 19443739
63. M. F. Berger, M. L. Bulyk, Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* **4**, 393–411 (2009). doi: 10.1038/nprot.2008.195; pmid: 19265799
64. M. T. Weirauch *et al*., Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014). doi: 10.1016/j.cell.2014.08.009; pmid: 25215497
65. N. Shen *et al*., Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding. *Cell Syst.* **6**, 470–483.e8 (2018). doi: 10.1016/j.cels.2018.02.009; pmid: 29605182
66. T. Siggers, M. H. Duyzend, J. Reddy, S. Khan, M. L. Bulyk, Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 555 (2011). doi: 10.1038/msb.2011.89; pmid: 22146299
67. T. Siggers, J. Reddy, B. Barron, M. L. Bulyk, Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding. *Mol. Cell* **55**, 640–648 (2014). doi: 10.1016/j.molcel.2014.06.019; pmid: 25042805
68. Y. Zhang, T. D. Ho, N. E. Buchler, R. Gordân, Competition for DNA binding between paralogous transcription factors determines their genomic occupancy and regulatory functions. *Genome Res.* **31**, 1216–1229 (2021). doi: 10.1101/gr.275145.120; pmid: 33975875
69. A. Isakova *et al*., SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods* **14**, 316–322 (2017). doi: 10.1038/nmeth.4143; pmid: 28092692
70. Y. Yin *et al*., Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017). doi: 10.1126/science.aaj2239; pmid: 28473536
71. P. Klein, T. Pawson, M. Tyers, Mathematical modeling suggests cooperative interactions between a disordered polyvalent ligand and a single receptor site. *Curr. Biol.* **13**, 1669–1678 (2003). doi: 10.1016/j.cub.2003.09.027; pmid: 14521832
72. S. Brodsky *et al*., Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol. Cell* **79**, 459–471.e4 (2020). doi: 10.1016/j.molcel.2020.05.032; pmid: 32553192
73. Y. Katan-Khaykovich, Y. Shaul, Nuclear import and DNA-binding activity of RFX1. Evidence for an autoinhibitory mechanism. *Eur. J. Biochem.* **268**, 3108–3116 (2001). doi: 10.1046/j.1432-1327.2001.02211.x; pmid: 11358531
74. A. S. Krois, H. J. Dyson, P. E. Wright, Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11302–E11310 (2018). doi: 10.1073/pnas.1814051115; pmid: 30420502
75. X. Wang *et al*., Dynamic Autoinhibition of the HMGB1 Protein via Electrostatic Fuzzy Interactions of Intrinsically Disordered Regions. *J. Mol. Biol.* **433**, 167122 (2021). doi: 10.1016/j.jmb.2021.167122; pmid: 34181980
76. S. Schütz *et al*., The Disordered MAX N-terminus Modulates DNA Binding of the Transcription Factor MYC:MAX. *J. Mol. Biol.* **434**, 167833 (2022). doi: 10.1016/j.jmb.2022.167833; pmid: 36174765
77. X. Wang *et al*., Negatively charged, intrinsically disordered regions can accelerate target search by DNA-binding proteins. *Nucleic Acids Res.* **51**, 4701–4712 (2023). doi: 10.1093/nar/gkad045; pmid: 36774964
78. T. Shimizu *et al*., Crystal structure of PHO4 bHLH domain-DNA complex: Flanking base recognition. *EMBO J.* **16**, 4689–4697 (1997). doi: 10.1093/emboj/16.15.4689; pmid: 9303313
79. E. Marklund *et al*., Sequence specificity in DNA binding is mainly governed by association. *Science* **375**, 442–445 (2022). doi: 10.1126/science.abg7427; pmid: 35084952
80. S. M. Sawaya, A. T. Bagshaw, E. Buschiazzo, N. J. Gemmell, Promoter microsatellites as modulators of human gene expression. *Adv. Exp. Med. Biol.* **769**, 41–54 (2012). doi: 10.1007/978-1-4614-5434-2_4; pmid: 23560304
81. S. Meinhardt, M. W. Manley Jr., D. J. Parente, L. Swint-Kruse, Rheostats and toggle switches for modulating protein function. *PLOS ONE* **8**, e83502 (2013). doi: 10.1371/journal.pone.0083502; pmid: 24386217
82. L. Swint-Kruse, C. Larson, B. M. Pettitt, K. S. Matthews, Fine-tuning function: Correlation of hinge domain interactions with functional distinctions between LacI and PurR. *Protein Sci.* **11**, 778–794 (2002). doi: 10.1110/ps.4050102; pmid: 11910022
83. M. D. Ilsley *et al*., Krüppel-like factors compete for promoters and enhancers to fine-tune transcription. *Nucleic Acids Res.* **45**, 6572–6588 (2017). doi: 10.1093/nar/gkx441; pmid: 28541545
84. B. Ho, A. Baryshnikova, G. W. Brown, Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome. *Cell Syst.* **6**, 192–205.e3 (2018). doi: 10.1016/j.cels.2017.12.004; pmid: 29361465
85. M. Castellanos, N. Mothi, V. Muñoz, Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nat. Commun.* **11**, 540 (2020). doi: 10.1038/s41467-019-14217-8; pmid: 31992709
86. A. M. Alexandari *et al*., De novo inference of thermodynamic binding energies using deep learning models of in vivo transcription factor binding. bioRxiv (2023); .doi: 10.1101/2023.05.11.540401
87. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
88. C. A. Davis *et al*., The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46** (D1), D794–D801 (2018). doi: 10.1093/nar/gkx1081; pmid: 29126249
89. M. B. Gerstein *et al*., Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012). doi: 10.1038/nature11245; pmid: 22955619
90. S. M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Curran Associates, Inc., 2017), vol. 30; https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
91. A. Shrikumar, P. Greenside, A. Kundaje, "Learning Important Features Through Propagating Activation Differences" in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 2017), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153; https://proceedings.mlr.press/v70/shrikumar17a.html.
92. M. A. Hume, L. A. Barrera, S. S. Gisselbrecht, M. L. Bulyk, UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **43**, D117–D122 (2015). doi: 10.1093/nar/gku1045; pmid: 25378322
93. S. A. Lambert *et al*., The human transcription factors. *Cell* **172**, 650–665 (2018). doi: 10.1016/j.cell.2018.01.029; pmid: 29425488
94. T. Gera, F. Jonas, R. More, N. Barkai, Evolution of binding preferences among whole-genome duplicated transcription factors. *eLife* **11**, e73225 (2022). doi: 10.7554/eLife.73225; pmid: 35404235
95. C. A. Shively, J. Liu, X. Chen, K. Loell, R. D. Mitra, Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16143–16152 (2019). doi: 10.1073/pnas.1818015116; pmid: 31341088
96. S. Feng *et al*., Transcription factor paralogs orchestrate alternative gene regulatory networks by context-dependent cooperation with multiple cofactors. *Nat. Commun.* **13**, 3808 (2022). doi: 10.1038/s41467-022-31501-2; pmid: 35778382
97. X. Zhou, E. K. O'Shea, Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol. Cell* **42**, 826–836 (2011). doi: 10.1016/j.molcel.2011.05.025; pmid: 21700227
98. D. D. Le *et al*., Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3702–E3711 (2018). doi: 10.1073/pnas.1715888115; pmid: 29588420
99. M. Levo *et al*., Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* **25**, 1018–1029 (2015). doi: 10.1101/gr.185033.114; pmid: 25762553
100. T. Gao, J. Qian, EnhancerAtlas 2.0: An updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–D64 (2020). pmid: 31740966
101. R. Gemayel, M. D. Vinces, M. Legendre, K. J. Verstrepen, Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010). doi: 10.1146/annurev-genet-072610-155046 pmid: 20809801
102. E. Segal, J. Widom, From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat. Rev. Genet.* **10**, 443–456 (2009). doi: 10.1038/nrg2591; pmid: 19506578
103. P. Korber, T. Luckenbach, D. Blaschke, W. Hörz, Evidence for histone eviction in *trans* upon induction of the yeast *PHO5* promoter. *Mol. Cell. Biol.* **24**, 10965–10974 (2004). doi: 10.1128/MCB.24.24.10965-10974.2004; pmid: 15572697
104. A. Valouev *et al*., A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008). doi: 10.1101/gr.076463.108; pmid: 18477713
105. S. L. Klemm, Z. Shipony, W. J. Greenleaf, Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019). doi: 10.1038/s41576-018-0089-8; pmid: 30675018
106. S. Bensmihen *et al*., The homologous ABI5 and EEL transcription factors function antagonistically to fine-tune gene expression during late embryogenesis. *Plant Cell* **14**, 1391–1403 (2002). doi: 10.1105/tpc.000869; pmid: 12084834
107. A. Rizzino, Transcription factors that behave as master regulators during mammalian embryogenesis function as molecular rheostats. *Biochem. J.* **411**, e5–e7 (2008). doi: 10.1042/BJ20080479; pmid: 18363551
108. J. F. Kribelbauer, C. Rastogi, H. J. Bussemaker, R. S. Mann, Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu. Rev. Cell Dev. Biol.* **35**, 357–379 (2019). doi: 10.1146/annurev-cellbio-100617-062719; pmid: 31283382
109. A. K. Csink, S. Henikoff, Something from nothing: The evolution and utility of satellite repeats. *Trends Genet.* **14**, 200–204 (1998). doi: 10.1016/S0168-9525(98)01444-9; pmid: 9613205
110. O. G. Berg, R. B. Winter, P. H. von Hippel, Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20**, 6929–6948 (1981). doi: 10.1021/bi00527a028; pmid: 7317363
111. T. Paramanathan, D. Reeves, L. J. Friedman, J. Kondev, J. Gelles, A general mechanism for competitor-induced dissociation of molecular complexes. *Nat. Commun.* **5**, 5207 (2014). doi: 10.1038/ncomms6207; pmid: 25342513
112. T. Jana, S. Brodsky, N. Barkai, Speed-specificity trade-offs in the transcription factors search for their genomic binding sites. *Trends Genet.* **37**, 421–432 (2021). doi: 10.1016/j.tig.2020.12.001; pmid: 33414013
113. J. Iwahara, Y. Levy, Speed-stability paradox in DNA-scanning by zinc-finger proteins. *Transcription* **4**, 58–61 (2013). doi: 10.4161/trns.23584; pmid: 23412360
114. I. Dror, R. Rohs, Y. Mandel-Gutfreund, How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays* **38**, 605–612 (2016). doi: 10.1002/bies.201600005; pmid: 27192961
115. J. V. W. Meeussen *et al*., Transcription factor clusters enable target search but do not contribute to target gene activation. *Nucleic Acids Res.* **51**, 5449–5468 (2023). doi: 10.1093/nar/gkad227; pmid: 36987884

116. S. Redding, E. C. Greene, How do proteins locate specific targets in DNA? *Chem. Phys. Lett.* **570**, 1–11 (2013). doi: 10.1016/j.cplett.2013.03.035; pmid: 24187380

117. A. B. Kolomeisky, Physics of protein-DNA interactions: Mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.* **13**, 2088–2095 (2011). doi: 10.1039/C0CP01966F; pmid: 21113556

118. A. Bhattacherjee, Y. Levy, Search by proteins for their DNA target site: 1. The effect of DNA conformation on protein sliding. *Nucleic Acids Res.* **42**, 12404–12414 (2014). doi: 10.1093/nar/gku932; pmid: 25324308

119. M. Bauer, R. Metzler, In vivo facilitated diffusion model. *PLOS ONE* **8**, e53956 (2013). doi: 10.1371/journal.pone.0053956; pmid: 23349772

120. A. Erbaş, J. F. Marko, How do DNA-bound proteins leave their binding sites? The role of facilitated dissociation. *Curr. Opin. Chem. Biol.* **53**, 118–124 (2019). doi: 10.1016/j.cbpa.2019.08.007; pmid: 31586479

121. R. I. Kamar *et al*., Facilitated dissociation of transcription factors from single DNA binding sites. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E3251–E3257 (2017). doi: 10.1073/pnas.1701884114; pmid: 28364020

122. L. M. Tuttle *et al*., Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. *Cell Rep.* **22**, 3251–3264 (2018). doi: 10.1016/j.celrep.2018.02.097; pmid: 29562181

123. K. Shrinivas *et al*., Enhancer Features that Drive Formation of Transcriptional Condensates. *Mol. Cell* **75**, 549–561.e7 (2019). doi: 10.1016/j.molcel.2019.07.009; pmid: 31398323

124. A. L. Sanborn *et al*., Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife* **10**, e68068 (2021). doi: 10.7554/eLife.68068; pmid: 33904398

125. H. Li, V. Rhodius, C. Gross, E. D. Siggia, Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11772–11777 (2002). doi: 10.1073/pnas.112341999; pmid: 12181488

126. J. J. Ferrie, J. P. Karr, R. Tjian, X. Darzacq, "Structure"-function relationships in eukaryotic transcription factors: The role of intrinsically disordered regions in gene regulation. *Mol. Cell* **82**, 3970–3984 (2022). doi: 10.1016/j.molcel.2022.09.021; pmid: 36265487

127. M. Slattery *et al*., Absence of a simple code: How transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014). doi: 10.1016/j.tibs.2014.07.002; pmid: 25129887

128. Ž. Avsec *et al*., The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019). doi: 10.1038/s41587-019-0140-0; pmid: 31138913

129. T. Kaplan *et al*., Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLOS Genet.* **7**, e1001290 (2011). doi: 10.1371/journal.pgen.1001290; pmid: 21304941

130. A. Tanay, Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006). doi: 10.1101/gr.5113606; pmid: 16809671

131. E. Morgunova *et al*., Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife* **7**, e32963 (2018). doi: 10.7554/eLife.32963; pmid: 29638214

132. J. M. Rogers *et al*., Bispecific Forkhead Transcription Factor FoxN3 Recognizes Two Distinct Motifs with Different DNA Shapes. *Mol. Cell* **74**, 245–253.e6 (2019). doi: 10.1016/j.molcel.2019.01.019; pmid: 30826165

133. L. Zhang *et al*., SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. *Genome Res.* **28**, 111–121 (2018). doi: 10.1101/gr.222844.117; pmid: 29196557

134. M. Franchina *et al*., The CD30 gene promoter microsatellite binds transcription factor Yin Yang 1 (YY1) and shows genetic instability in anaplastic large cell lymphoma. *J. Pathol.* **214**, 65–74 (2008). doi: 10.1002/path.2258; pmid: 17973241

135. J. Yao *et al*., Transcription factor ICBP90 regulates the MIF promoter and immune susceptibility locus. *J. Clin. Invest.* **126**, 732–744 (2016). doi: 10.1172/JCI81937; pmid: 26752645

136. V. Boeva *et al*., De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.* **38**, e126–e126 (2010). doi: 10.1093/nar/gkq217; pmid: 20375099

137. S. R. Yant *et al*., High affinity YY1 binding motifs: Identification of two core types (ACAT and CCAT) and distribution of potential binding sites within the human β globin cluster. *Nucleic Acids Res.* **23**, 4353–4362 (1995). doi: 10.1093/nar/23.21.4353; pmid: 7501456

138. R. Hopfner *et al*., Genomic structure and chromosomal mapping of the gene coding for ICBP90, a protein involved in the regulation of the topoisomerase IIalpha gene expression. *Gene* **266**, 15–23 (2001). doi: 10.1016/S0378-1119(01)00371-7; pmid: 11290415

139. D. E. Riley, J. N. Krieger, Short tandem repeat (STR) replacements in UTRs and introns suggest an important role for certain STRs in gene expression and disease. *Gene* **344**, 203–211 (2005). doi: 10.1016/j.gene.2004.09.034; pmid: 15656986

140. K. Sathasivam *et al*., Aberrant splicing of *HTT* generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2366–2370 (2013). doi: 10.1073/pnas.1221891110; pmid: 23341618

141. M. Baralle, T. Pastor, E. Bussani, F. Pagani, Influence of Friedreich ataxia GAA noncoding repeat expansions on pre-mRNA processing. *Am. J. Hum. Genet.* **83**, 77–88 (2008). doi: 10.1016/j.ajhg.2008.06.018; pmid: 18597733

142. A. A. Shishkin *et al*., Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. *Mol. Cell* **35**, 82–92 (2009). doi: 10.1016/j.molcel.2009.06.017; pmid: 19595718

143. J. Hui, K. Stangl, W. S. Lane, A. Bindereif, HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat. Struct. Biol.* **10**, 33–37 (2003). doi: 10.1038/nsb875; pmid: 12447348

144. Ł. J. Sznajder, M. S. Swanson, Short Tandem Repeat Expansions and RNA-Mediated Pathogenesis in Myotonic Dystrophy. *Int. J. Mol. Sci.* **20**, 3365 (2019). doi: 10.3390/ijms20133365; pmid: 31323950

145. Ł. J. Sznajder *et al*., Intron retention induced by microsatellite expansions as a disease biomarker. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4234–4239 (2018). doi: 10.1073/pnas.1716617115; pmid: 29610297

146. C. A. Nutter *et al*., Cell-type-specific dysregulation of RNA alternative splicing in short tandem repeat mouse knockin models of myotonic dystrophy. *Genes Dev.* **33**, 1635–1640 (2019). doi: 10.1101/gad.328963.119; pmid: 31624084

147. C. S. Shelley, F. E. Baralle, Deletion analysis of a unique 3′ splice site indicates that alternating guanine and thymine residues represent an efficient splicing signal. *Nucleic Acids Res.* **15**, 3787–3799 (1987). doi: 10.1093/nar/15.9.3787; pmid: 3108860

148. H. Cuppens *et al*., Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes. The polymorphic (Tg)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation. *J. Clin. Invest.* **101**, 487–496 (1998). doi: 10.1172/JCI639; pmid: 9435322

149. J. Hui *et al*., Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. *EMBO J.* **24**, 1988–1998 (2005). doi: 10.1038/sj.emboj.7600677; pmid: 15889141

150. K. Hamanaka *et al*., Genome-wide identification of tandem repeats associated with splicing variation across 49 tissues in humans. *Genome Res.* **33**, 435–447 (2023). doi: 10.1101/gr.277335.122; pmid: 37307504

151. K. Yap *et al*., A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. *Mol. Cell* **72**, 525–540.e13 (2018). doi: 10.1016/j.molcel.2018.08.041; pmid: 30318443

152. G. V. Echeverria, T. A. Cooper, RNA-binding proteins in microsatellite expansion disorders: Mediators of RNA toxicity. *Brain Res.* **1462**, 100–111 (2012). doi: 10.1016/j.brainres.2012.02.030; pmid: 22405728

153. K. Ninomiya, T. Hirose, Short tandem repeat-enriched architectural RNAs in nuclear bodies: functions and associated diseases. *Noncoding RNA* **6**, 6 (2020). doi: 10.3390/ncrna6010006; pmid: 32093161

154. M. Lynch *et al*., A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9272–9277 (2008). doi: 10.1073/pnas.0803466105; pmid: 18583475

155. C. E. Pearson, K. Nichol Edamura, J. D. Cleary, Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.* **6**, 729–742 (2005). doi: 10.1038/nrg1689; pmid: 16205713

156. G. Levinson, G. A. Gutman, Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221 (1987). pmid: 3328815

157. G. F. Richard, F. Pâques, Mini- and microsatellite expansions: The recombination connection. *EMBO Rep.* **1**, 122–126 (2000). doi: 10.1093/embo-reports/kvd031; pmid: 11265750

158. I. Tirosh, N. Barkai, K. J. Verstrepen, Promoter architecture and the evolvability of gene expression. *J. Biol.* **8**, 95 (2009). doi: 10.1186/jbiol204; pmid: 20017897

159. E. K. Farley *et al*., Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015). doi: 10.1126/science.aac6948; pmid: 26472909

160. C. J. Markin *et al*., Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, eabf8761 (2021). doi: 10.1126/science.abf8761; pmid: 34437092

161. R. K. Das, S. L. Crick, R. V. Pappu, N-terminal segments modulate the α-helical propensities of the intrinsically disordered basic regions of bZIP proteins. *J. Mol. Biol.* **416**, 287–299 (2012). doi: 10.1016/j.jmb.2011.12.043; pmid: 22226835

162. N. Lyle, R. K. Das, R. V. Pappu, A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.* **139**, 121907 (2013). doi: 10.1063/1.4812791; pmid: 24089719

163. C. G. Kalodimos *et al*., Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **305**, 386–389 (2004). doi: 10.1126/science.1097064; pmid: 15256668

164. I. Cortes-Ciriano, S. Lee, W.-Y. Park, T.-M. Kim, P. J. Park, A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 15180 (2017). doi: 10.1038/ncomms15180; pmid: 28585546

165. R. Wooster *et al*., Instability of short tandem repeats (microsatellites) in human cancers. *Nat. Genet.* **6**, 152–156 (1994). doi: 10.1038/ng0294-152; pmid: 8162069

166. K. Brower *et al*., An open-source, programmable pneumatic setup for operation and automated control of single- and multi-layer microfluidic devices. *HardwareX* **3**, 117–134 (2018). doi: 10.1016/j.ohx.2017.10.001; pmid: 30221210

167. F. Chollet, "Keras" (GitHub, 2018); https://github.com/fchollet/keras.

168. M. Abadi *et al*., TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467v2 [cs.DC] (2016).

169. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG] (2017).

170. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). doi: 10.1093/nar/27.2.573; pmid: 9862982

171. R. Andersson *et al*., An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014). doi: 10.1038/nature12787; pmid: 24670763

172. J. Wang *et al*., HACER: An atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019). doi: 10.1093/nar/gky864; pmid: 30247654

173. S. Ossowski *et al*., The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**, 92–94 (2010). doi: 10.1126/science.1180677; pmid: 20044577

174. T. N. Marriage *et al*., Direct estimation of the mutation rate at dinucleotide microsatellite loci in Arabidopsis thaliana (Brassicaceae). *Heredity* **103**, 310–317 (2009). doi: 10.1038/hdy.2009.67; pmid: 19513093

175. J. M. Watson *et al*., Germline replications and somatic mutation accumulation are independent of vegetative life span in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12226–12231 (2016). doi: 10.1073/pnas.1609686113; pmid: 27729523

176. C. A. Horton *et al*., Code for: Short tandem repeats bind transcription factors to tune eukaryotic gene expression, Zenodo (2023); doi: 10.5281/zenodo.8161422

ACKNOWLEDGMENTS

We thank C. Markin, N. DelRosso, S. Thompson, D. Le, D. Miller, S. Levy, D. Herschlag, J. Rutledge, R. Lee, K. Collins, and members of the Fordyce laboratory for technical assistance, valuable discussions, and feedback on the manuscript. **Funding:** This work was supported by the National Science Foundation (grant CAREER-2142336 to P.M.F. and grant MCB-1715589 to R.G.), the National Institutes of Health (grant R01-GM117106 to R.G. and grant DP2-GM123641 to P.M.F.), and the Swedish Research Council (grant 2020-06459 to E.M.). P.M.F. is a Chan Zuckerberg Biohub Investigator. N.S. and J.Z. were supported by the Stowers Institute for Medical Research. P.H.S. was supported by the Stanford Bio-X Bowes Fellowship. A.K.A. was supported by the ChEM-H Chemistry-Biology Interface predoctoral training program. J.M.S., A.K.A., and M.G.B.H. were supported by the National Science Foundation Graduate Research Fellowships Program. **Author contributions:** C.A.H. and M.G.B.H. measured transcription factor-DNA affinities and kinetics in MITOMI microfluidic assays. C.A.H. performed EMSA assays. A.K.A. performed STAMMP assays for the Pho4 mutant library. A.M.A., A.S., and A.K. developed, trained, and tested NN models on genomic binding data. C.A.H. and J.M.S. designed and implemented the Gillespie algorithm simulation of TF binding kinetics. E.M. derived and fit microscopic rate model. P.H.S. designed the microfluidic devices. C.A.H., A.K.A., P.H.S., and M.G.B.H. fabricated microfluidic molds and devices. C.A.H. performed bioinformatic analyses and statistical mechanical modeling. N.S. and J.Z. guided the development of NN models and analysis of chromatin immunoprecipitation data. C.A.H., A.A., R.G., J.Z., A.K., W.J.G., and P.M.F. conceived of the study. C.A.H. and P.M.F. wrote the manuscript with input from all authors. P.M.F. supervised the project. **Competing interests:** A.K. is on the scientific advisory board of PatchBio, SerImmune, AINovo, TensorBio, and OpenTargets. A.K. was a consultant with Illumina Inc. until Jan 2023. A.K. owns stock/shares in DeepGenomics, Immunai, Freenome, PatchBio and SerImmune. P.M.F. is on the scientific advisory board of Evozyme. The remaining authors declare no competing interests. **Data and materials availability:** Data not presented in the main text or supplementary materials are available at Zenodo (*55*). Code to reproduce analyses is available at Zenodo (*176*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add1250
Materials and Methods
Supplementary Text
Figs. S1 to S156
Tables S1 to S16
References (*177–192*)
MDAR Reproducibility Checklist

Submitted 24 May 2022; resubmitted 12 May 2023
Accepted 26 July 2023
10.1126/science.add1250