

Reviewed Preprint

v1 • February 24, 2025

Not revised

Reviewed Preprint

v2 • April 9, 2026

Revised by authors

For correspondence:

andreas.wagner@ieu.uzh.ch

Competing interests:

No competing interests declared

Funding:

See page 64

Reviewing editor: Vincent J Lynch, University at Buffalo, State University of New York, United States

© 2025, Antunes Westmann et al.

This article is distributed under the terms of the [Creative Commons](#)

[Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

The adaptive landscapes of three global *Escherichia coli* transcriptional regulators

Cauã Antunes Westmann^{1,2}, Leander Goldbach^{1,2}, Andreas Wagner^{1,2,3}✉

¹Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland • ²Swiss Institute of Bioinformatics, Lausanne, Switzerland • ³The Santa Fe Institute, Santa Fe, United States

eLife Assessment

This study maps the genotype-phenotype landscapes of three *E. coli* transcription factors and the topographical features of these landscapes. It shows that ruggedness and epistasis do not hinder the evolution of strong transcription factor binding sites. These **convincing** findings contribute **important** insights into fitness landscape theories and highlight the role of chance, contingency, and evolutionary biases in gene regulation. The authors then study the topographical features of these landscapes, especially the number and distribution of local maxima, as well as the statistical properties of evolutionary paths on these landscapes.

<https://doi.org/10.7554/eLife.103774.2.sa3>

Abstract

The evolution of gene regulation is a major source of evolutionary adaptation and innovation, particularly when organisms encounter new or changing environments. Central to this process is the emergence of new transcription factor binding sites (TFBSs). Adaptive landscapes provide a powerful framework to study such emergence by linking regulatory DNA sequences to their transcriptional outputs. Although several landscapes have been characterized for DNA, RNA, and proteins, large-scale in vivo adaptive landscapes for bacterial TFBSs remain scarce. Here, we address this gap by experimentally mapping the first comprehensive in vivo regulatory landscapes for three global transcription factors in *Escherichia coli*: CRP, Fis, and IHF. Using a massively parallel reporter assay, we quantify the regulation strength of more than 30,000 TFBS variants for each factor and reconstruct their adaptive landscapes. All three landscapes are highly rugged and exhibit pervasive epistasis, with thousands of local peaks distributed broadly across sequence space. This ruggedness contrasts sharply with the much smoother TFBS landscapes of eukaryotes. It suggests greater constraints on the evolution of prokaryotic gene regulation. Nonetheless, evolutionary simulations show that ~10% of evolving populations can reach a peak of strong regulation, a proportion that is significantly greater than in comparable random landscapes. Adaptive evolution starting from the same DNA sequence can attain different high peaks, and some peaks are reached more frequently than others. Together, our results show that de novo adaptive evolution of new gene regulation in bacteria is feasible, but subject to a blend of chance, historical contingency, and evolutionary biases.

Introduction

Transcriptional regulation controls the ability of RNA polymerase to initiate the transcription of a gene into mRNA¹. It is crucial in the life of all organisms, orchestrating the expression of genes in response to environmental cues and cellular states^{2,3}. Transcriptional regulation is mediated by transcription factors (TFs), proteins that bind DNA near a gene at short DNA words known as transcription factor binding sites (TFBSs). The binding of a TF to its binding site on DNA can either

hinder or facilitate transcription initiation by RNA polymerase^{1,4,5}. The stronger the binding of a TF to its TFBS is, the more strongly the TF can regulate a nearby gene^{6–9}. In bacteria, TFs operate within a hierarchically organized gene regulatory network. At the bottom of this hierarchy are *local* TFs that regulate the expression of one or few genes and modulate specific biological processes^{10–12}. At the top are *global* TFs that may regulate hundreds of genes and control many cellular functions^{11,13–15}. Consistent with their broader role, global regulators are typically more highly expressed and bind to a broader range of TFBSs than local regulators¹².

A genotype-phenotype map is a conceptual analog of a physical landscape, where each location corresponds to one genotype in sequence space, and is associated with a quantitative phenotype. If the phenotype is related to gene expression, the map is also called a regulatory landscape^{6,16–19}. Another special case of such a map is a fitness landscape or adaptive landscape, in which the phenotype is fitness and is interpreted as an elevation^{20,21}. Over the last decade, genotype-phenotype (GP) maps and fitness landscapes have become central tools for understanding how molecular systems evolve under mutation and selection^{22–25}. Such maps and landscapes have been experimentally studied for DNA^{6,8,18,19,26,27}, protein^{28–32} and RNA^{33–35} molecules, revealing key topographical properties that shape evolutionary outcomes, including epistasis^{24,36}—the non-additive effects of multiple mutations on phenotype—landscape ruggedness, reflected in the number and distribution of fitness peaks, and constraints on adaptive evolution. For example, one large-scale study adopted CRISPR-Cas9 technology to measure the fitness of more than 200,000 *E. coli* genotypes that encode variants of the bacterial antibiotic resistance gene dihydrofolate reductase (DHFR). It showed that this landscape is highly rugged (multi-peaked)³¹.

For transcription factor binding sites, most pertinent large-scale studies are based on in vitro binding assays, such as protein-binding microarrays (PBMs), and they focus predominantly on eukaryotic transcription factors⁶. While these studies have been instrumental in characterizing transcription factor binding preferences, they typically do not measure regulatory output in a native cellular context. In contrast, comprehensive in vivo data for bacterial TFBSs remain extremely rare. To our knowledge, only two high-resolution in vivo landscapes have been previously mapped for bacterial regulators, those of the local regulators TetR¹⁸ and LacI²⁷. As a result, it remains unclear whether principles inferred from protein landscapes, eukaryotic TFBSs, or in vitro binding assays generalize to transcriptional regulation in bacteria, particularly for global regulators¹¹ that integrate multiple physiological signals.

Both TFs and their TFBSs evolve, but TFBSs evolve more rapidly. The reason is that a mutation in a TF can affect the expression of many genes, whereas a mutation in a TFBS may affect only one gene and is thus less likely to be deleterious^{13,37–39}. A special case of TFBS evolution is the evolution of a strong TFBS from a DNA word with weak or no regulatory activity. Such *de novo* evolution of TFBSs can create new regulatory interactions and change the structure of gene regulatory networks^{40–43}. Unfortunately, we know little about the ability of Darwinian evolution to create TFBSs *de novo*. A strong TFBS may have to arise from a non-binding site through an evolutionary path of multiple mutational steps. Darwinian evolution can favor this process only if strong binding is adaptive and if each mutational step in a path increases binding strength, i.e., if the path is *accessible* to Darwinian evolution. To find whether such paths exist may require the analysis of multiple paths. This is challenging because sequence space contains an astronomical number of potential TFBSs for any one TF. The number of evolutionary paths to strong transcriptional regulation is thus also astronomical. For each such path, the strength of each TFBS along the path has to be measured experimentally^{44–47}.

In principle, one could attempt to construct such landscapes in silico using commonly employed models of TF–DNA interactions, such as position weight matrices (PWMs)^{48–50}. However, PWMs are derived from a limited set of naturally occurring binding sites and primarily reflect sequence conservation rather than quantitative regulatory output. Moreover, PWMs assume independent and additive contributions of individual nucleotide positions to DNA binding^{48–50}. They therefore cannot capture the influences of epistatic interactions between positions, which can dramatically affect landscape topography and evolutionary accessibility²⁴. Lastly, PWMs do not account for important biological effects that modulate gene regulation such as DNA shape^{51,52}, cooperative

interactions^{53,54}, and chromosomal context^{55,56}. Thus, experiments are necessary not only to obtain quantitative measurements of gene regulatory activity, but also to refine and inform PWM-based models using thousands of experimentally characterized sequences.

Building on our previous work on a local TF¹⁸, we address this challenge for three global regulators in *Escherichia coli* by performing three independent and massively parallel experiments⁴⁴ for each TF. The experiments use a synthetic biology platform to quantify how strongly each of more than 30,000 binding sites for a TF can regulate the expression of a nearby reporter gene.

The first TF we study is the cAMP receptor protein (CRP). In the absence of glucose, CRP modulates the expression of genes mostly involved in carbon metabolism. It allows *E. coli* to efficiently switch between sources of carbon and energy^{57–60}. The second TF is the factor for inversion stimulation (Fis), which helps to regulate the expression of genes involved in growth phase transitions. It also modulates the supercoiling of DNA^{61–63} and influences DNA replication, recombination, and repair^{61–63}. The third factor is the integration host factor (IHF). IHF regulates genes involved in stress responses and stationary phase survival^{64,65}. Similar to Fis, it is involved in DNA compaction, replication, and recombination, but also in the assembly of complex nucleoprotein structures^{64,65}. We chose these factors because they are the most global regulators in *E. coli*, and they are diverse, belonging to different protein families.

We use our experimental data for each TF's binding sites to map the relationship between binding site genotype and gene expression. Our first aim is to characterize the resulting regulatory landscapes for global bacterial regulators, and to find out whether these landscapes are different or similar. When strong regulation of a gene is adaptive, a regulatory landscape becomes a fitness landscape^{20,21,25,66}. Our second aim is to understand how populations would evolve on each of our three landscapes when they are viewed as fitness landscapes. Specifically, we study how evolving populations would traverse each landscape through individual mutational steps that change a TFBS its ability to regulate a gene via its cognate TF. A peak in such a landscape is a TFBS conveying stronger regulation than all neighboring TFBSs in sequence space. If such a landscape is rugged (has multiple peaks), natural selection alone may not enable a population to discover the highest peaks, i.e., the strongest TFBSs. The reason is that the peaks may be separated by valleys of low regulation strength that cannot be traversed by natural selection alone^{67–69,25}. In other words, high peaks may not be easily accessible through Darwinian evolution – they may be reachable by few or no evolutionary paths of single DNA mutations in which each mutational step increases TFBS strength⁶. One factor that can reduce peak accessibility is epistasis, which can reduce the predictability of evolutionary trajectories towards a peak^{24,36}.

To accomplish both aims, we first studied the topography of the three landscapes, and then simulated the dynamics of evolving populations on them. All three landscapes are highly rugged. They contain more than 2,000 peaks that are scattered through genotype space and are rife with epistatic interactions, in striking contrast to the comparatively smooth TFBS landscapes described for eukaryotic systems⁶. Despite these features, evolving populations can reach the strongest TFBSs in all three landscapes. Which of several high peak is reached is contingent on chance events during adaptive evolution.

Results

Landscape mapping

We constructed a modular plasmid system based on the common backbone plasmid pCAW-Sort-Seq-V2¹⁸ (Supplementary Table S1 [↗](#); Supplementary Figures S1 [↗](#)–S2 [↗](#); **Supplementary Methods 2–3**). This backbone contains all shared regulatory and reporter elements required for fluorescence-based measurements, but it encodes neither a TF nor its binding site(s) for any of the regulators studied here. From this backbone, we generated three TF-specific plasmid derivatives. Each of these plasmids encodes one of the global transcription factors CRP, Fis, or IHF under inducible control (hereafter pCAW-Sort-Seq-V2-CRP, pCAW-Sort-Seq-V2-Fis, and pCAW-Sort-Seq-V2-IHF; Supplementary Figure S2 [↗](#)). In each of the TF-specific plasmids, a TFBS insertion site is

positioned immediately upstream of the *gfp* reporter gene, such that TF binding represses the transcription of *gfp*. Consequently, stronger TF–DNA binding results in lower GFP expression, which enables a quantitative readout of a binding site’s regulation strength.

For each TF-specific plasmid, we constructed three kinds of variants. The first carries a wild-type (WT) TFBS for the corresponding TF upstream of *gfp*. It serves as a reference conferring wild-type regulation of the reporter gene. The second contains a complete TFBS library, in which we randomized the eight most information-rich base-pair positions of the respective binding site (**Supplementary Methods 4**; [Supplementary Table S3](#)), as determined from alignments of experimentally characterized binding sites curated in RegulonDB⁷⁰. Each library comprised $4^8 = 65,536$ unique TFBS sequences, and we constructed three independent biological replicate plasmid libraries from them per TF (**Supplementary Methods 4–5**). The third variant lacks a promoter upstream of *gfp* and serves as a negative control that allows us to quantify cellular autofluorescence during fluorescence measurements ([Supplementary Table S1](#)).

We introduced each TF-specific plasmid into an *E. coli* host strain in which the corresponding chromosomal TF gene had been deleted (Δcrp , Δfis , or Δihf ; [Supplementary Table S2](#)). This genetic background ensures that the focal transcription factor is expressed exclusively from the plasmid. Although the mutant strains grow more slowly than the WT, they reached similar cell densities during late exponential or early stationary phase, the growth phase at which we performed all measurements ([Supplementary Figure S3](#)). TF expression is controlled by a tetracycline-inducible promoter and can be precisely tuned using anhydrotetracycline (anTc), allowing us to regulate TF abundance independently of growth conditions and to isolate the effects of TF–DNA binding on transcriptional regulation ([Figure 1a](#)).

We then mapped these TFBS genotypes to their respective regulatory phenotypes using a well-established technique known as sort-seq^{16,44,71–73} ([Figure 1b](#)), which combines fluorescence-activated cell sorting (FACS) with high-throughput sequencing ([Figure 1c](#)). In sort-seq, one first sorts cells into multiple fluorescence “bins” depending on the level of GFP expression. A cell’s GFP fluorescence serves as a proxy for GFP expression levels⁷⁴ and quantifies how strongly the TFBS library member in this cell can regulate GFP expression (**Methods, Supplementary Methods 5**). For each TFBS genotype, we quantified regulation strength (S) as a weighted average of its sequencing counts across the different fluorescence bins, yielding a single continuous measure of regulatory activity (see **Methods**).

The results of our sort-seq experiments are three maps – one for each transcription factor – from each of more than 30,000 genotypes (TFBS variants) to regulatory phenotypes (regulation strength). One can view each map as a regulatory landscape^{8,9,16,18,75–79} ([Figure 1d](#)) that becomes a fitness landscape whenever strong regulation entails high fitness^{6,25}. For the purpose of analyzing the landscape quantitatively, we represent it as a network of genotypes (TFBS variants). Each node in this network corresponds to a genotype. Edges link neighboring genotypes, which differ in a single nucleotide ([Figure 1d](#)).

Landscapes exhibit diverse regulation strengths and distribution breadths

To evaluate the ability of our library to regulate gene expression, we first measured the distribution of GFP fluorescence intensities across the bins in two conditions, i.e., in the presence or absence of the TF (with or without the atc inducer, [Supplementary figures S4–S6](#)). In the presence of the TF, GFP expression was lower on average and showed a broader distribution than in the absence of the TF ([Supplementary figures S4–S6](#)). This indicates that the TFBSs in each library can indeed downregulate GFP expression, but some library variants convey stronger regulation than others, hence the broader fluorescence distribution.

We then pooled barcoded DNA sequences extracted from cells in each bin and biological replicate. We sequenced at least 250 unique TFBS genotypes from each bin, with an average of 3992.3 ± 4293.9 unique sequences per bin for the three TFs (as detailed in **Methods, Supplementary Methods 7**, and [Supplementary Figures S4–S6](#)). The resulting sequences covered 95%, 90%, and 93% of

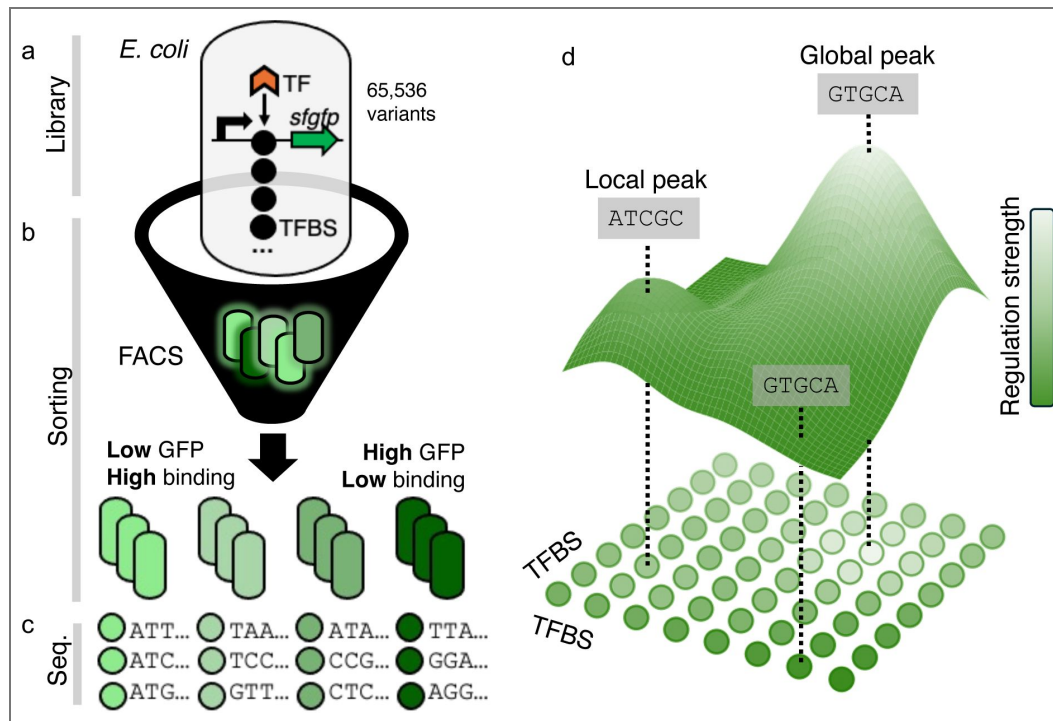


Figure 1. Mapping TFBS regulatory landscapes.

a-c) Sort-Seq procedure. We utilized a plasmid-based fluorescence reporter system followed by sort-seq to map TFBS regulatory landscapes. **a. Library generation.** For generating our *E. coli* libraries, we cloned TFBS sequence variants ($4^8 = 65,536$), represented as black circles, into our plasmid, between a σ^{70} constitutive promoter (shown as a black right-facing arrow) and a *gfp* gene (shown as a green right-facing arrow), to measure transcriptional repression through fluorescence intensity. When a TF binds to the TFBS, it blocks RNA polymerase activity through steric hindrance, reducing *gfp* transcription. Thus, lower fluorescence levels (light green) represent stronger TF-TFBS binding. The TFBS variants produce a range of regulation strengths resulting in variable green fluorescence intensities in bacterial cells (green-colored rounded rectangles). **b. Sorting procedure.** We sorted libraries into expression bins based on fluorescence intensity using a fluorescence-activated cell sorting (FACS, Supplementary Figures S4–S6). **c. Sequencing and phenotyping.** We sequenced TFBS variants from each fluorescence bin and used these data to calculate a continuous regulation strength for each genotype. Regulation strength (*S*) was computed as a weighted average of fluorescence across bins, based on the distribution of sequencing reads (see Methods). Regulation strength is visualized using a color gradient from dark green (low-affinity TFBSs, high fluorescence) to light green (high-affinity TFBSs, low fluorescence). **d) Genotype-phenotype mapping.** To construct a regulatory landscape, we connected TFBS genotypes (colored circles) that differed by a single nucleotide via edges, thereby establishing an interconnected genotype network. Each genotype conveys a regulatory phenotype (strength of regulation, heatmap colors), which can be viewed as the elevation dimension (z-axis) in a landscape.

the total library sizes ($N = 65,536$ genotypes) for CRP, Fis and IHF, respectively. To ensure the reliability of our data, we excluded sequences with low read coverage and sequences that were not present in all triplicates (**Methods, Supplementary Methods 7.3**). This quality filtering step resulted in library sizes of 31,975 genotypes for CRP (49% of all 4^8 genotypes), 43,222 genotypes for Fis (66%), and 41,325 genotypes for IHF (63%). The correlation in read counts for each variant across replicates was very high for this quality-filtered data (Pearson's $R = 0.98-0.99$) (Supplementary Figures S7–S9). We used this data for all further analyses.

The fluorescence and sequence data from each bin allowed us to quantify the variant's ability to regulate gene expression for each TFBS library variant. We refer to the resulting metric as the *regulation strength* S conveyed by the variant (**Methods, Supplementary Methods 7.2, Figure 1d**). It ranges between $S=0$ (no regulation) to $S=1$ (strongest regulation among all variants in the library). Although our experiments directly quantify expression regulation, and not the affinity or binding strength of a TF to a TFBS variant, we also use *binding strength* as a proxy for regulation strength, because TF-DNA binding is necessary for regulation^{1,4,11,54,80,81}. To each of our three TFs, we also assigned a reference TFBS that is naturally occurring and conveys strong regulation by the TF, as proven by previous experimental work^{60,82–84}. We refer to this TFBS as the wild-type (WT^{82,85} WT^{83,86} and WT^{84,87}, Supplementary Table S4). We refer to TFBSs with regulation strengths below and above the wild-type (WT_{CRP}: $S_{WT}=0.71$, WT_{Fis}: $S_{WT}=0.97$, WT_{IHF}: $S_{WT}=0.95$) as weak and strong, respectively.

We observed a broad range of regulation strengths S for each TF landscape, with varying dispersions. The CRP landscape exhibited the lowest average regulation strength and a narrower distribution of S compared to Fis and IHF, which had similar distributions (mean $\pm$ s.d., 0.37 ± 0.1 for CRP, 0.57 ± 0.13 for Fis, and 0.52 ± 0.14 for IHF; see Figure 2a, Supplementary Figure S10). Next, we analyzed the strongly regulating TFBSs to identify nucleotides that may be particularly frequent, and thus potentially important for strong regulation (Figure 2b). We discovered a moderate association between the most frequent nucleotides (highlighted in yellow in Figure 2b) and the most informative nucleotides from the available position weight matrices (PWMs) for these TFs⁷⁰ (Pearson correlation coefficients: $R=0.51$, $R=0.43$, $R=0.47$ for Fis, CRP and IHF, respectively, with p -values smaller than 10^{-16} , rejecting the null hypothesis of an absence of association). This observation suggests that our logos capture different information compared to available PWMs. This is expected, because our approach allows us to filter sequences by regulation strength thresholds to construct our matrices, unlike the traditional method of aligning genomic TFBSs without considering their regulation strengths^{48,49}.

To further validate the data from our sort-seq experiments, we isolated cells harboring 10 different TFBS variants from each of 13 bins of each library (i.e., $10\times13=130$ variants per library), and determined their regulation strength more directly by quantifying GFP expression with a microplate reader (**Methods** and Supplementary Figure S11). This comparison validates the sort seq approach by revealing a strong association of the two independent quantifications of regulation strength (Pearson's $R=0.81$ for CRP, $R=0.74$ for Fis, and $R=0.73$ for IHF). (**Methods** and Supplementary Figure S11).

All three regulatory landscapes are highly rugged

The study of our landscapes in a network framework (Figure 1d) can help to quantify different aspects of landscape topography (Supplementary Table S5). One of them is the ruggedness of each landscape. It can be quantified by the number of peaks^{31,88–90}. In the network framework, a peak is a TFBS whose neighbors are all weaker regulators (with lower S) than the TFBS itself. We find that all three landscapes are highly rugged, with 2,154, 2,312, and 2,453 peaks for the CRP, Fis, and IHF landscapes, respectively (Figure 2c, Supplementary Figure S12, Supplementary Table S5). Not surprisingly, peak genotypes generally are stronger regulators than non-peak genotypes (Figure 2c). Only a small fraction of peak genotypes regulate expression more strongly than the wild-type (Figure 2c; 61 for CRP, 172 for Fis, and 199 for IHF). We refer to such peaks as high peaks and distinguish them from low peaks ($S < S_{WT}$).

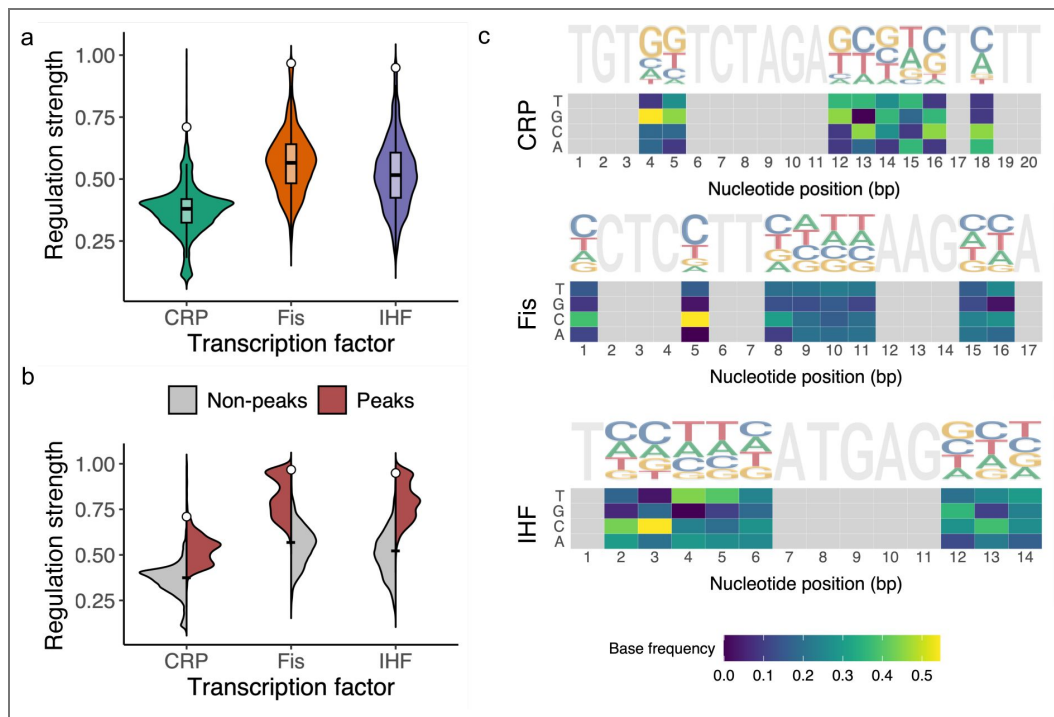


Figure 2.

a. Genotypes in each landscape vary broadly in their regulation strength. The violin plots show the distribution of regulation strengths S (vertical axis) for each TF landscape (horizontal axis). The width of a plot at a given value of S represents the frequency of TFBSs at this value. The vertical length of the box in each violin plot covers the range between the first and third quartiles (IQR). The horizontal line within the box represents the median value, and whiskers span 1.5 times the IQR. The white circle shows the regulation strength of the wild-type for each landscape (CRP: 0.71, Fis: 0.97, IHF: 0.95). The landscape sizes are as follows. CRP: $N=31,975$ TFBS variants; Fis: $N=43,222$; IHF: $N=41,325$. **b. Peak genotypes are stronger regulators than non-peak genotypes.** Dual violin plots show the distribution of regulation strength (vertical axis) for the three TF landscapes (horizontal axis), stratified by non-peak genotypes (grey, CRP, $N=29,821$, Fis, $N=40,910$, IHF, $N=38,872$), and peak genotypes (red, CRP, $N=2,154$, Fis, $N=2,312$, IHF, $N=2,453$). The black tick-mark in each plot indicates the mean regulation strength of both non-peak and peak genotypes taken together (mean $\pm$ standard deviation, CRP: 0.37 ± 0.1 , Fis: 0.57 ± 0.13 , IHF: 0.52 ± 0.14). The white circle on each plot marks the regulation strength of the wild-type (CRP: 0.71, Fis: 0.97, IHF: 0.95). **c. Sequence logos and nucleotide frequency matrices for strong CRP, Fis and IHF binding sites.** Each sequence logo ^{48,49} is based on an alignment of TFBSs with greater regulation strength than the wild-type for each of the three TFs IHF, Fis, and CRP. Each logo also shows the non-varying position (grey) of the TFBS genotype from which our libraries were created. In each logo, the height of each letter at each TFBS position indicates the information content at that nucleotide position – the taller the letter, the more frequent the nucleotide is in strongly regulating TFBSs ^{48,49}. Similar information is conveyed by the frequency matrices displayed as heat maps below each logo. They represent the variability of each nucleotide at each position (horizontal axis) through a color gradient (see color legend). Tall letters in the sequence logo and yellow letters in the frequency matrix indicate frequent, and thus likely important nucleotides for strong regulation.

The prevalence of sign epistasis (Supplementary Table S5 [↗](#)) supports the notion that our landscapes are indeed rugged (see **Supplementary Methods 7.5** for further details on epistasis and its evolutionary consequences). Independent evidence for landscape ruggedness comes from comparing the ruggedness of our landscapes with that of a well-established theoretical model of uncorrelated random landscapes, in which each sequence is assigned a fitness at random from the same fitness distribution, and neighbors have uncorrelated fitness values^{90–92}. Such landscapes are maximally rugged^{90–92}. We created 10^3 uncorrelated random landscapes for each of our three TF landscapes by randomly shuffling the measured regulation strengths among all genotypes (**Supplementary Methods 7.6**). The number of peaks in our TF landscapes lies within 93%, 96%, and 98% percent of that of a maximally rugged random landscape, which, on average [mean±s.d.], has $2,308\pm133$, $2,405\pm89$ and $2,373\pm97$ peaks for CRP, Fis and IHF, respectively. This analysis underlines that our landscapes are indeed highly rugged.

Because we use only quality-filtered genotype data, our landscapes lack regulatory information for about 40% of the 4^8 genotypes. While we cannot exclude that this undersampling of genotypes has led to systematic biases in our determination of landscape ruggedness, we note that the sampling of landscape genotypes by our experiments was not strongly biased with respect to regulation strength. Specifically, when we analyzed the relative connectivity of genotypes in our landscapes – the fraction of each genotype's 24 possible neighbors for which our experiments yielded regulatory data – we found that it is only weakly correlated with regulation strength ($R=-0.1$, -0.1 , 0.01 for the CRP, Fis, and IHF landscapes, [Figures S13–S15](#) [↗](#)). Similarly, the relative connectivity of peak genotypes is only weakly correlated with their regulation strength ($R=-0.05$, -0.04 , 0.06 for the CRP, Fis, and IHF landscapes).

Landscape peaks are widely scattered in genotype space

In a rugged adaptive landscape, reaching high fitness peaks can be challenging, because such peaks are separated from other genotypes by valleys of low fitness that cannot be traversed by natural selection alone²⁵. If natural selection favors strong gene regulation, the ruggedness of our regulatory landscapes may thus present a challenge for adaptive evolution. To better understand the potential magnitude of this challenge, we next analyzed our landscapes' topography in greater detail. We began by studying the distribution of peaks in genotype space. If a landscape's highest peaks are widely scattered through genotype space, then they may be accessible via fitness-increasing evolutionary paths from diverse non-peak genotypes. This may facilitate adaptive evolution compared to a landscape where peaks are clustered in a small region of genotype space.

For each of our three landscapes, we determined the genetic distance between peaks, i.e., the minimum number of mutations needed to transition from one peak to another, regardless of their effect on regulation strength. We compared the distribution of these distances to the distribution of genetic distances for an equal number of randomly selected non-peak variants. In all three landscapes, the distances between peaks are almost indistinguishable from those of random genotypes, differing on average by fewer than 0.1 mutational steps. In other words, peaks are about as widely dispersed in each landscape as random genotypes ([Supplementary Figure S16](#) [↗](#)). This is the case for both low and high peaks ([Supplementary Figures S17–S18](#) [↗](#)). A principal component analysis further underscores this dispersion (**Supplementary Methods 7.7**, [Supplementary Figures S19–S21](#) [↗](#)).

Accessible paths to a peak are often not the shortest possible paths

Next, we focused on mutational paths to peaks that are evolutionarily accessible, i.e., a series of mutational steps where each step increases regulation strength. Specifically, we enumerated all accessible paths that exist from each non-peak genotype to each peak genotype for all three landscapes. We found that these paths are generally longer than the shortest genetic distance between a non-peak and the peak genotypes. Specifically, the mean length of accessible paths

exceeded the shortest distance by 1.5, 1.8 and 1.8 steps for the CRP, Fis and IHF landscapes (Supplementary Figure S22 [↗](#)). In other words, accessible mutational paths are often not the most direct paths.

The existence of indirect paths implies that some mutational steps are evolutionarily prohibited because they decrease regulation strength. The reason is closely linked to non-additive (epistatic) effects of two or more mutations on regulation strength. (see **Supplementary Methods 7.5** for further details on epistasis). More specifically, such inaccessible steps are a result of sign epistasis. In this kind of epistasis, a double mutant of a TFBS regulates expression more strongly than the TFBS itself, even though one or both constituent single mutants regulate expression more weakly than the TFBS^{24,93–95}. Indeed, epistasis is prevalent in all three landscapes (Table S5 [↗](#)). Specifically, we observe sign epistasis in 62%, 66%, and 65% percent of interactions between single mutant pairs in the CRP, Fis, and IHF landscapes.

The existence of accessible paths alone does not tell us how easily a high peak can be found through Darwinian evolution. The reason is that only a very small fraction of all paths to that peak may be accessible, and an evolving population may not find any one of those paths. To quantify the fraction of accessible paths, we determined the total number of paths from each non-peak genotype to each high fitness peak, and computed the fraction of those paths that are accessible. Figure 3a [↗](#) shows how this fraction depends on the number of mutational steps in a path. It behaves similarly for all three landscapes (Figure 3a [↗](#)). The majority of two-step paths (a fraction greater than 80%) are accessible in all three landscapes (Figure 3a [↗](#)), but the fraction of accessible paths dwindles rapidly with an increasing number of mutational steps. It reaches a minimum below 1% for all three landscapes at eight mutational steps (Figure 3a [↗](#)).

High peaks have large basins of attraction that share many TFBS genotypes

In the following analysis, we computed the basin of attraction of each peak, defined as the set of non-peak genotypes from which accessible paths to the peak exist^{31,96,97}. In other words, a peak's basin of attraction comprises all genotypes from which adaptive evolution can access the peak. The peaks in our landscapes have widely different basin sizes. They include peaks accessible from a mere fraction of genotypes to those accessible by a majority of genotypes (Figure 3b [↗](#)). Notably, high peaks generally had larger basins of attraction than low peaks (Figure 3c [↗](#), Supplementary Figure S23 [↗](#)). In addition, we found that many genotypes are part of multiple basins of attraction. Adaptive evolution may reach different high peaks starting from any such genotype. For all pairs of high peaks, we computed the pairwise overlap (intersection) of the basins of attraction, i.e., the fraction of genotypes that are part of both basins of attraction. (Figure 3d [↗](#), Supplementary Figures S24 [↗](#)–S25 [↗](#)). The number of genotypes in this intersection varies widely, but the basins of attraction of high peaks generally share a substantial fraction of genotypes (mean shared genotypes: 48%, 42%, and 36% among all pairs of high peaks for the CRP, Fis, and IHF landscapes, Figure 3d [↗](#)).

Genetic drift facilitates and clonal interference reduces the attainment of high peaks

Taken together, our analyses thus far indicate that high peaks in the CRP, Fis, and IHF landscapes are more accessible than low peaks (Figure 3b [↗](#)), and their basins of attraction also share more genotypes (Figure 3d [↗](#)). However, the accessible evolutionary paths to high peaks are often indirect. In addition, the fraction of accessible paths to any one high peak dwindles rapidly with the distance from that peak (Figure 3a [↗](#)).

To understand how these topographical features jointly affect adaptive evolution, we simulated the evolutionary dynamics on our three landscapes under the assumption that natural selection would favor strong regulation and that fitness is proportional to regulation strength.

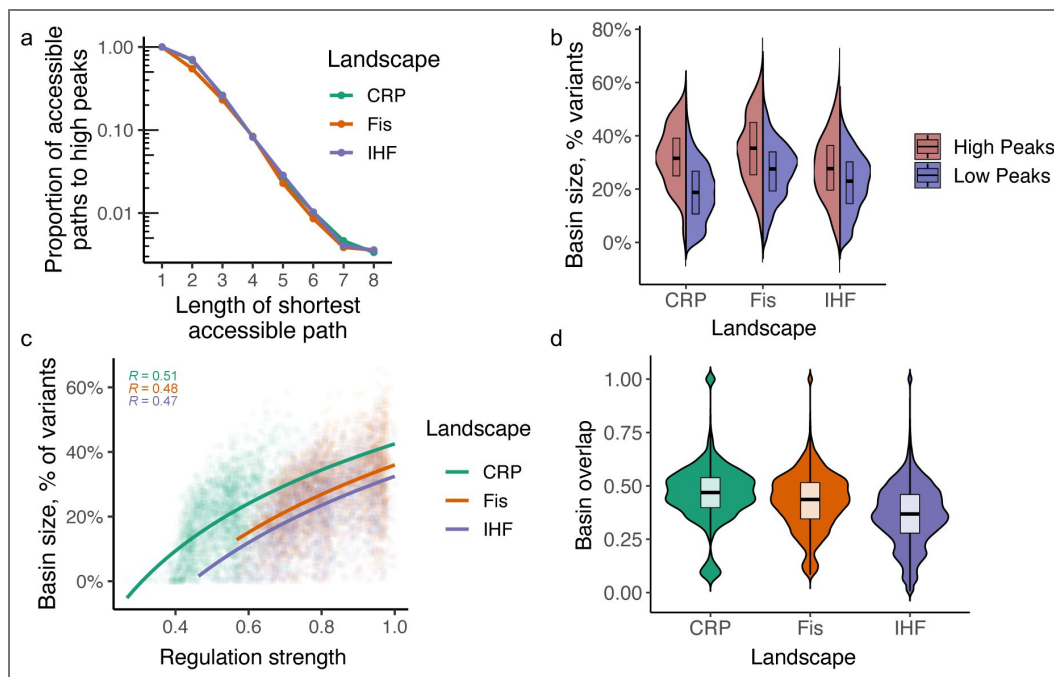


Figure 3. Peaks and their basins of attractions.

a. The fraction of accessible paths declines with increasing path length. The figure illustrates how the fraction of accessible paths to high peaks (vertical axis, logarithmic scale) decreases with the length of the shortest accessible path (horizontal axis). Accessible paths are defined as paths where each step increases regulation, starting from a specified initial genotype. Each colored line corresponds to data from a different TF landscape: CRP (green), Fis (orange), and IHF (blue). Circles indicate the mean fraction of accessible paths for a given path length. **b. Higher peaks have larger basins of attraction.** The split-violin plots display the distribution of basin sizes (vertical axis) for high (red) and low (blue) peaks in the three TF landscapes CRP, Fis, and IHF (horizontal axis). High peaks have significantly larger basins of attraction (Welch Two Sample t-tests; CRP: t-value = 9.0898, df = 63.509, and p-value = 4.22×10^{-13} , with mean basin sizes of $x_{\text{high}} = 9528.426$ and $x_{\text{low}} = 5655.035$. Fis: t-value = 7.617, df = 172.35, and p-value = 1.645×10^{-12} , with mean basin sizes of $x_{\text{high}} = 14206.70$ and $x_{\text{low}} = 10931.94$. IHF: t-value = 6.1777, df = 201.99, and p-value = 3.521×10^{-9} , with mean basin sizes of $x_{\text{high}} = 10965.872$ and $x_{\text{low}} = 8664.953$). The violin plots show the distribution of basin sizes for each of the two kinds of peaks. Their width represents the frequency of a given basin size. The vertical length of the box in each plot covers the range between the first and third quartiles (IQR). The horizontal line within the box represents the median value, and whiskers span 1.5 times the IQR. **c. Peak genotypes with larger basins of attraction regulate expression more strongly.** The regulation strength of peak genotypes (horizontal axis) is plotted against the size of their basins of attraction (vertical axis, shown as a percentage of all (non-peak genotypes) for peaks in all three landscapes (color legend). Color-coded values of R at the top of the graph indicate Pearson correlation coefficients for each landscape (color legend). Curved lines are based on a linear regression analysis (in linear space), with grey shading indicating 95% confidence intervals (CRP: $R^2=0.28$, N = 2,154; Fis: $R^2=0.24$, N = 2,312; IHF: $R^2=0.23$, N = 2,434). **d. Basins of attractions share many TFBS genotypes.** The violin plots with embedded boxplots illustrate the fraction of genotypes shared between basins of attractions (vertical axis) for all pairs of high peaks in the CRP (green), Fis (orange), and IHF (purple) TF landscapes (horizontal axis). Specifically, we computed basin overlap as one minus the Jaccard coefficient (Supplementary Methods 7.5). A value of one (zero) indicates that the basin of attraction of two peaks share all (no) genotypes (CRP: 61 high peaks, N=1,830 comparisons, Fis: 172 high peaks, N=14,706 comparisons, IHF: 199 high peaks, N=19,701 comparisons). Their width represents the frequency of a given basin overlap. The vertical length of the box in each box covers the range between the first and third quartiles (IQR). The horizontal line within the box represents the median value, and whiskers span 1.5 times the IQR.

Because high peaks constitute only a small fraction of all peaks (2.8%, 0.4%, and 0.5% in the CRP, Fis, and IHF landscapes), we hypothesized that most evolving populations would reach only low peaks. To test this hypothesis, we took advantage of the fact that *E. coli* has a low genomic mutation rate of 2.2×10^{-10} per base pair per generation⁶⁷, and that we study evolution only within an 8-nucleotide segment of a TFBS. In addition, *E. coli* has a large effective population size ($>10^8$ individuals⁶⁷), which means that genetic drift is weak and even minor differences in fitness are visible to natural selection^{67,98}. These conditions imply that a population on our landscape would evolve in the well-studied strong-selection weak-mutation regime (SSWM)^{16,99–101}. In this regime, a population is monomorphic most of the time, until a beneficial mutation arises, which usually sweeps rapidly to fixation. In other words, adaptive evolution can be modeled as an adaptive walk, in which each mutational step is taken with a fixation probability that has been derived by Kimura⁹⁸. We thus call the resulting adaptive walks Kimura walks (**Supplementary Methods 8**). We simulated 10^3 such adaptive walks starting from each of 15,000 randomly and uniformly selected starting (non-peak) TFBS genotypes from each landscape. Each random walk comprised up to 25 mutational fixation steps, unless it reached a fitness peak earlier. Each adaptive walk also accounted for known *E. coli* mutational biases^{102–104} (**Supplementary Methods 8**).

As we hypothesized, most adaptive walks indeed reach only low peaks (90%, 85%, and 87% percent for the CRP, Fis, and IHF landscapes.). However, the percentage of walks that terminate at high peaks is several-fold greater than the proportion of high peaks. Specifically, 10 percent of walks terminate at high peaks in the CRP landscape, which is 3.6-fold higher than the percentage (2.8%) of high peaks in this landscape. In the Fis and IHF landscape, adaptive walks are 2 and 1.6-fold more likely to terminate at high peaks than expected from the proportion of these peaks among all peaks (7.4% and 8.1% percent) (Figure 4a⁴).

To determine whether high regulatory peaks are more accessible due to chance alone, we compared the empirical landscapes to randomized “null” landscapes. Specifically, we generated these randomized landscapes by permuting regulation strengths across genotypes while preserving the sampled genotype network (**Supplementary Methods 7.6**). On each randomized landscape, we then performed adaptive-walks (15,000 walks for each of 10^3 starting genotypes) with the same parameters as for the biological landscape. For all three landscapes, the fraction of adaptive walks reaching high regulatory peaks in the empirical landscapes exceeds that for the randomized landscapes by almost three-fold, a difference that is statistically significant (Supplementary Figure S30⁴, **Supplementary Methods 9**). In sum, rugged regulatory landscapes strongly constrain evolutionary trajectories, yet do not render the evolution of strong regulation vanishingly rare. Instead, strong regulatory phenotypes remain evolutionarily attainable at levels that exceed null expectations, even though they are reached by only a minority of evolving populations.

The adaptive walks that reached a high peak were short (Figure 4b⁴). On average, they were also no more than half a mutational step longer than the shortest genetic distance between each starting genotype and peak (Figure 4b⁴, CRP landscape: path length 3.2 ± 1.6 [mean $\pm$ s.d.] vs. genetic distance 2.8 ± 1.2 ; Fis landscape: path length 3.1 ± 1.5 vs. genetic distance 2.7 ± 1.2 ; IHF landscape: path length 3.1 ± 1.5 vs. genetic distance 2.7 ± 1.2).

These short paths can be explained by our previous observation that peaks are widely distributed across the genotype spaces of the three landscapes. This distribution makes it easier for populations starting from any (non-peak) genotype to find a local peak through few mutations. Additionally, path lengths realized during adaptive walks tend to be shorter than the average lengths of accessible paths. (Figure 4b⁴).

Because genetic drift can cause evolving populations to escape a low fitness peak and attain a higher fitness peak, we also asked how small population sizes affect the likelihood that a population attains a high fitness peak (**Supplementary Material 8**, Supplementary Figures S26⁴–S28⁴). When simulating adaptive evolution in populations of $N=10^2$ individuals, we found that the likelihood that a population reaches a high peak increased by at least 10% percent (to 18%, 20%, and 21% for the CRP, Fis, and IHF landscapes, Supplementary Figures S26⁴–S28⁴).

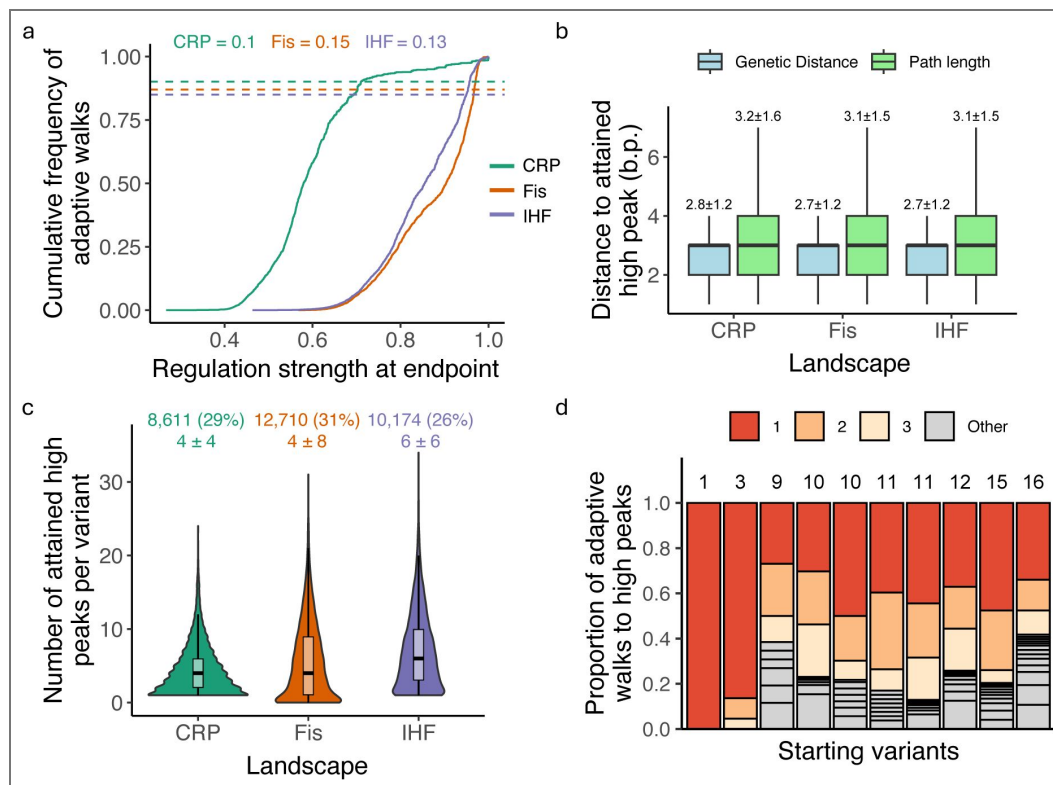


Figure 4. Peak accessibility, contingency and biases in adaptive walks.

a. More than 10% percent of adaptive walks lead to high peaks. The graph shows the cumulative distribution function (CDF, vertical axis) of regulation strength (horizontal axis) attained at the end of 15 million adaptive walks (15,000 starting genotypes $\times$ 1,000 adaptive walks each) for each TF landscape (color legend, CRP: green; Fis: orange; IHF: purple). Dashed lines intersect the CDF at points equivalent to the wild-type regulation strength for each TF (color legend). They help to infer the proportion y_{TF} of walks that terminate at high peaks ($S > S_{WT}$), which is indicated by the numerical values at the graph's top for each of the landscapes (color legend). **b. Paths to high peaks are not much longer than genetic distances from the starting genotype.** Colored boxplots summarize, for each of the three landscapes (horizontal axis), the distribution of shortest genetic distances (blue) and actual path lengths (green) for adaptive walks terminating at high peaks. Path lengths are typically less than one mutational step longer than genetic distances. Each box covers the range between the first and third quartiles (IQR). The horizontal line within each box represents the median value, and whiskers span 1.5 times the IQR. Numbers atop each plot indicate means $\pm$ one standard deviation. **c. Different numbers of high peaks are reached from different starting variants.** Violin plots integrated with boxplots show the distribution in the number of high peaks reached (vertical axis) from each starting variant that attained any high peak during 1,000 adaptive walks.

We also studied how deviations from the strong selection weak mutation regime would affect evolutionary dynamics. In large populations or at high mutation rates, populations tend to be polymorphic and are subject to clonal interference, where the most advantageous of several co-occurring mutations typically dominates and becomes fixed^{105–107}. To approximate this dynamic, we conducted simulations using “greedy” adaptive walks (**Supplementary Methods 8**)^{25,108}. In such an adaptive walk, it is always a genotype’s mutational neighbor with the highest increase in regulation strength that becomes fixed^{100,108–110}. In other words, a greedy walk starting from a given genotype is deterministic. We, therefore, simulated only one walk for each of the 15,000 randomly chosen (non-peak) starting genotypes. We found that the fraction of greedy walks reaching high regulatory peaks is slightly lower than in the SSWM regime, with 1%, 2%, and 5% fewer walks achieving such peaks in the CRP, Fis, and IHF landscapes, respectively (**Supplementary Figure S29**). This outcome is expected, because greedy walks prioritize immediate fitness gains at the expense of the ability to discover higher fitness peaks¹⁰⁸.

The attainment of peaks is highly contingent on chance events

Because different basins of attraction share many TFBS genotypes (**Figure 3d**, **Supplementary Figures S24–S25**), it is possible that adaptive evolution starting from any one genotype can reach different peaks, depending on which mutational paths it takes. In other words, the structure of the landscapes we study may lead to evolutionary contingency^{111,112} – the dependence of an outcome of evolution on unpredictable chance events^{47,111,113}. Indeed, non-peak genotypes can access on average around 30% of the total number of high peaks in each landscape (**Supplementary Figure S25**). To assess the prevalence of evolutionary contingency, we determined how many different peaks are attained by 10^3 adaptive walks starting from the same randomly chosen genotype. We performed this analysis on 15,000 starting genotypes, focusing on the subset of those starting genotypes from which at least one high peak is reached during the 10^3 walks. Specifically, 28.6%, 23%, and 24.5% of starting genotypes for the CRP, Fis, and IHF landscapes, reached at least one high peak. Notably, these percentages represent the fraction of starting genotypes that can access at least one high peak, not the total fraction of adaptive walks that lead to a high peak, which is lower (as shown in **Figure 4a**). For instance, while adaptive walks starting from 28.6% of genotypes reached at least one high peak in the CRP landscape only 10% of all adaptive walks in this landscape end at a high peak (**Figure 4a**). Importantly, for any one starting genotype where random walks reached at least one high peak, they typically reached more than one high peak (**Figure 4c**, median±IQR, 4±4, 4±8, 6±6 for the CRP, Fis, and IHF landscapes, respectively). The number of high peaks attained from any one starting genotype was as large as 21, 30, and 33 for the three respective landscapes. Thus, evolutionary contingency is pervasive in all three landscapes. In addition, not only the end point of evolution but also the evolutionary trajectories leading to this end point show contingency. That is, adaptive walks starting from a specific genotype and ending at a specific high peak often take multiple paths to this peak (**Supplementary Figure S31**).

Biases towards specific evolutionary outcomes have been documented in both empirical and computational studies^{18,31,47,111,114–116}. They also exist in our landscapes (**Figure 4d**, **Supplementary Figure S31**). **Figure 4d** illustrates both contingency and bias for 10^3 adaptive walks starting from each of 10 different genotypes in the CRP landscape (thus a total of 10^4 adaptive walks). Depending on the starting variant, the adaptive walks reached between 1 and 16 high peaks. Whenever different peaks are reached, they are reached by markedly different proportions of adaptive walks. The same holds in the Fis and IHF landscapes (**Supplementary Figure S33**). Additionally, while multiple routes to each peak can be traversed (**Supplementary Figure S31**), we observed in all three landscapes biases towards specific paths that are taken more frequently than others (**Supplementary Figure S34**).

The number of these variants is shown as an integer on top of the panel, and its percentage among all 15,000 starting variants is shown in parentheses (color legend). Underneath it, the panel shows the mean ± 1 s.d. of the number of attained high peaks. **a. d. Some high peaks are reached more often than others.** We randomly and uniformly sampled 10 starting genotypes from the CRP

landscape, started 10^3 adaptive walks from each, and recorded the number and frequency of distinct high peaks attained in these random walks. Results for each starting genotype are symbolized by a vertical bar. The number of stacks within each bar (delineated by horizontal lines, also indicated by an integer above each bar) indicates the number of high peaks reached by the 10^3 adaptive walks. Starting variants are ordered in ascending order based on this number of attained peaks. Stack height indicates the fraction of walks that reached the same peak, and is indicated in red, orange, and yellow for the three most frequently attained peaks. See [Supplementary Figure S33](#) for the Fis and IHF landscapes.

Lastly, because sort-seq measurements are subject to experimental uncertainty^{71,117,118}, an important question is whether such noise affects the inferred structure of a regulatory landscape and the resulting evolutionary dynamics. To address this issue we explicitly incorporated empirically estimated measurement uncertainty into our fitness comparisons and repeated pertinent analyses under two conditions: a noise-free landscape ($\mathcal{G}^S$) and an “uncertainty-aware” landscape $\mathcal{G}^{S,\tau}$ using genotype-specific noise estimates (τ , see [Supplementary Figure S35](#), **Supplementary Methods 10**). We examined whether noise-induced changes in landscape structure influence evolutionary trajectories. To do so, we compared adaptive walk dynamics between the two kinds of landscapes using identical population genetic parameters. Despite differences in peak counts and local topology, the overall pattern of genotype visitation during adaptive walks was highly similar between the two kinds landscapes. Specifically, the visitation frequency profiles of genotypes were strongly correlated between $\mathcal{G}^S$ and $\mathcal{G}^{S,\tau}$ landscapes (Spearman’s $\rho < 0.001$ shown in [Supplementary Figure S35](#)). This means that genotypes frequently accessed in the noise-free landscape remain frequently accessed in the noisy landscape.

Discussion

We evaluated the ability of three *E. coli* global regulators to control gene expression through each of more than 30,000 TFBSs for each regulator. To this end, we utilized a synthetic plasmid-based system that facilitates high-throughput fluorescence measurements. This system allowed us to quantify the regulation strength of individual TFBSs by measuring gene expression through GFP fluorescence⁷⁴. Additionally, the system insulates the control of transcription from any direct effects library sequences might have on transcription, such as the transcriptional and translational impacts of 5’ untranslated regions on gene expression^{119,120}. Recent studies have investigated large empirical datasets of cis-regulatory genotypes, examining the impact of sequence variation on gene regulation in both eukaryotes and prokaryotes^{6,16,72,81,121–124}. However, few studies have focused on these interactions from an evolutionary perspective and studied adaptive landscapes of gene regulation, as we do here^{6,16,123}.

We showed that the regulatory landscapes of all three TFs are highly rugged and have multiple peaks. The ruggedness of all three landscapes is also supported by the prevalence of epistasis between pairs of TFBS mutations ([Supplementary Table S5](#)). A particularly important form of epistasis is sign epistasis^{24,93,94}, because it can lead to multiple adaptive peaks^{24,93,94} (see **Supplementary Methods 7.5**). Our landscapes contain up to 65% of mutation pairs with sign epistasis, a value that is especially high compared to the almost exclusively additive interactions of mutations in eukaryotic TFs^{6,125}. However, the TFs we study are not exceptions among other prokaryotic TFs. Recent smaller-scale studies have shown that epistatic interactions are common in binding sites of local prokaryotic regulators, such as AraC (sign epistasis in more than 50% of 20 mutants¹²⁶) and Cl (sign epistasis in 85% of 113 mutants¹²⁷).

A possible reason for this greater incidence of epistasis lies in the nature of prokaryotic TFBSs. Specifically, prokaryotic TFBSs are at approximately 20bps twice as long as eukaryotic TFBSs^{80,128} and exhibit symmetries that reflect the dimeric state of their cognate TFs^{129–131}. These factors may increase the likelihood of intramolecular epistasis. Our observations raise important questions for future work, such as why the landscapes of prokaryotic TFBSs differ so dramatically from those of eukaryotic ones. And what do these differences imply for the evolutionary dynamics of gene regulation?

Despite the high ruggedness and pervasive epistasis of these landscapes, we found that a modest fraction of evolving populations can still access the highest regulatory peaks. Specifically, our evolutionary simulations show that 10% of populations with a size typical of *E. coli* reach one of the highest peaks. This percentage is significantly higher than in randomized landscapes (Supplementary Methods 9; Supplementary Figure S30 [↗](#)), which shows that the structure of our regulatory landscapes facilitates access to stronger regulation than expected by chance. We speculate that this property reflects the biological role of global regulators, which coordinate the expression of many target genes¹¹ and thus must operate across a wide spectrum of regulation strengths, while still permitting the evolution of strong regulation when necessary.

The clonal interference that may occur in even larger populations or in populations with high mutation rates reduces the accessibility of high peaks from 10% to 5% of evolving populations (Supplementary Figure S29 [↗](#)). Conversely, in small populations where genetic drift can help a population escape from a low peak, this percentage increases to 18%. Numbers like these render the de novo evolution of strong transcription factor binding sites plausible for our three global regulators. Once such binding sites have originated in one population, they can also spread to others through horizontal gene transfer^{132,133}.

In addition to increasing the ruggedness of a landscape, epistasis can also influence the evolution of populations on the landscape^{24,94,134}. The regulation strength of a TFBS is partially determined by the combined interactions of the nucleotides within the TFBS^{50,135–137}, which in turn affects TF-TFBS binding affinities. Different combinations of mutations can result in similar regulation strengths and create multiple evolutionary pathways to achieve optimal or near-optimal regulation^{6,138} (Supplementary Figure S31 [↗](#)). This can lead to overlapping basins of attraction among different peak TFBSs conveying strong regulation, such that even evolution starting from the same genotypes can take different paths and reach different peaks with similar regulation strengths. Indeed, we observed such overlapping basins in our landscapes (Figure 3d [↗](#) and Supplementary Figure S24 [↗](#)). Moreover, epistasis can create plateaus of regulation strength where various combinations of nucleotides yield TFBSs with intermediate regulation strengths. These plateaus further contribute to the overlapping basins of attraction we observe, and may serve as common evolutionary intermediates that multiple starting sequences can traverse on their way to different peaks.

In a landscape with substantial epistasis, the sequence of mutations that occurs in an evolving population can also render adaptive evolution highly contingent on this sequence^{111,112}. For example, some beneficial mutations within a TFBS might only be accessible after other specific mutations have occurred. Different starting sequences or early mutations can also change the spectrum of accessible mutations, leading to different peaks with similarly strong regulation^{111,112}. Indeed, we observed in all three landscapes that different evolving populations starting from the same genotypes in the landscape attain different peaks (Figures 3d [↗](#), 4d [↗](#) and Supplementary Figure S33 [↗](#))^{111,113,139,140}. Such contingency reduces the predictability of evolution^{47,111,141}.

Despite the prevalence of contingency in all three landscapes, evolving populations starting from the same genotype more often attain some peaks than others (Supplementary Figure S33 [↗](#)). Moreover, for each attained peak with multiple possible evolutionary routes, we also observed that some paths are more frequently transversed than their alternatives (Supplementary Figure S34 [↗](#)). That is, evolution is biased towards traversing some paths and attaining some peaks more often than others^{142,143}. These observations emphasize the complex interplay between chance, contingency, and evolutionary biases in shaping the outcomes of adaptive evolution^{47,111,113,115}.

The three TFs we studied here belong to different protein families, yet they have regulatory landscapes with similar topography. All three landscapes are highly rugged, highly epistatic, and harbor multiple peaks that are widely scattered, and this holds for peaks of low, intermediate, and high regulation strength. In addition, peak accessibility in all three landscapes increases with peak height (Supplementary Figure S32 [↗](#)). On the one hand, these commonalities may be caused by common biological or biochemical properties of the three TFs. For example, CRP has been suggested to possess nucleoid-associated protein properties similar to Fis and IHF, due to its ability

to bend and loop DNA¹⁴⁴. On the other hand, the commonalities may reflect general characteristics of global gene regulation. One of them is that global TFs often bind unspecifically to multiple TFBSs^{11,13}. Also, they may bind DNA with a broad range of different affinities centered around intermediate affinity, rather than bind few sites but very strongly¹³. This possibility is supported by comprehensive analyses of in vitro eukaryotic TF binding affinities for thousands of TFBS variants, showing essentially continuous binding affinity distributions for hundreds of sites⁶.

Despite broad similarities among the three landscapes we study, we also found some differences. Most notable is the narrower distribution of regulation strengths for the binding sites of CRP compared to those of Fis and IHF (Figure 2a [↗](#)). This is not unexpected, given that CRP binding sites are known for being quasi-symmetric and less degenerate than those of Fis and IHF^{57,60}. Previous studies have also shown that Fis and IHF binding sites are less conserved in their nucleotide sequence and more biased toward AT-rich content^{43,145,146}. Differences in the distribution of regulation strengths may also result from differences in functions among the three proteins. CRP's function is mostly restricted to gene regulation. As a result, it may have evolved to finely tune its regulatory output, resulting in a narrower distribution of regulation strengths. In contrast, Fis and IHF are involved in functions beyond gene regulation, such as DNA replication and genomic structural maintenance^{147,148}, which might impose additional constraints or require a broader range of DNA binding strengths.

One limitation of our work stems from our rigorous quality filtering of sort-seq data. As a result, we lack regulatory data for some 40% of the TFBSs in each landscape. This limited diversity of reliable data is a common feature in mutational library studies. It has several technical causes, such as biases in library synthesis¹⁴⁹, PCR amplification¹⁵⁰, cloning, and loss of sequence diversity after cell sorting¹²². In future work, this limitation could be overcome by a combination of strategies, such as subsampling complete genotype spaces or combining different molecular methods to overcome biases and diversity loss during PCR amplification^{151,152}, sorting^{71,117,153}, and high-throughput sequencing¹²³.

Importantly, although undersampling of genotype space is a limitation, standard approaches such as random subsampling or predictive modeling are not straightforward remedies. Several of our core analyses—including peak identification, quantification of epistasis, and assessment of evolutionary accessibility—rely on combinatorially complete local neighborhoods in genotype space. Random subsampling of genotypes would remove mutational neighbors, and thereby confound pairwise comparisons and the interpretation of landscape topology. Predictive modeling could be used to infer missing genotypes and reconstruct more complete landscapes¹⁵⁴, but it requires additional assumptions that introduce their own limitations. In addition, developing, validating, and benchmarking such models would be beyond the scope of this study, which is focused on empirical landscape mapping, and can serve as a starting point for future modeling work.

A second limitation comes from our use of the sort-seq method, which is best-suited for our work, because it allows the high-throughput measurement and sorting of millions of individual cells in a standardized and straightforward manner. However, the method's accuracy depends on the binning procedure. Other studies have used between two and several dozen bins, with or without unbiased sampling^{71,73,81,117,121,122,155}. Recommendations emerging from this work include sorting of cells into at least four logarithmically ($\log_2$) equally-spaced bins, each covering approximately 12.5-15% of the fluorescence distribution^{81,122,156}. We followed these recommendations, using 13 bins. In addition, we computed a weighted average to calculate expression values from sequences appearing in multiple bins, a straightforward method validated by robust studies in transcriptional regulation analysis^{16,72}. In addition, we validated individual regulation strengths with an independent method to demonstrate its reliability.

A third limitation of our study is that we only examined variation at eight positions for each TF. We selected these positions based on their importance for DNA binding and regulation by our TFs, as evidenced by their high information content. We cannot exclude the possibility that selecting a different set of positions could yield different landscape topographies. However, we speculate that less information-rich position would not reduce but rather increase the potential for the de novo

evolution of strong TFBSs. For example, they may facilitate landscape navigability through extradimensional bypasses^{96,157} or provide small-incremental changes in regulation strengths. These small changes help populations reach peaks via diminishing returns effects, meaning that as a population gets closer to a peak, each subsequent mutation contributes progressively smaller improvements to regulation strength^{158,159}. Investigating a wider array of positions and larger regulatory landscapes remains an important task for future work.

Fourth, we use a simplified empirical system for the study of gene regulation—a promoter followed by a single TFBS. Although this regulatory architecture exists for some genes¹⁶⁰, global regulators often form part of more complex, combinatorial architectures^{4,160,161} that are influenced by environmental factors^{2,162,163}, concentrations of active TFs^{164–166}, chromosome structure^{167–171}, methylation states¹⁷², and more. Studies on more complex promoter architectures may reveal regulatory landscapes with different topographies.

Lastly, we also analyzed the adaptive evolution of TFBSs in a simplified manner, performing data-driven evolutionary simulations rather than experimental evolution. Although many studies assume that regulation strength and fitness are correlated, this is not always the case. Low binding affinities can be adaptive during development¹⁷³, and many genes exhibit a nonlinear fitness-expression function¹⁷⁴ with a plateau of maximal fitness across a wide range of expression levels^{175–177}. For instance, most mutations and polymorphisms in the promoter of the yeast gene *TDH3* do not significantly affect fitness in a glucose-rich medium¹⁷⁶. For these reasons we focused our observations and interpretations on regulatory phenotypes (regulation strength). Our simplified approach allowed us to model population evolution in a large genotype space and avoid monitoring thousands of evolving regulatory sequences simultaneously in vivo^{23,178}. However, it cannot replace experimental evolution of TFBSs, which also remains an important challenge for future work.

Transcription factor binding sites are among the simplest units of biological organization. Our work provides the first large-scale analysis of the regulatory landscapes formed by such sites for global transcriptional regulators. It shows that strong binding sites can readily evolve de novo, even though prokaryotic transcription factor binding sites are much larger than their eukaryotic counterparts and have more rugged regulatory landscapes. In addition, the evolution of these simple sequences also displays phenomena that have been characterized in much more complex systems, such as evolutionary contingency and evolutionary biases.

Materials and methods

The Supplementary Materials contain extended details of experimental procedures and data analysis.

Strains and plasmids

Bacterial strains and plasmids used in this work are listed in Table S1¹⁷⁹. We obtained electrocompetent *E. coli* cells of strain SIG10-MAX[®] from Sigma Aldrich (CMC0004). We used this strain for molecular cloning and library generation due to its high transformation efficiency. The genotype of this strain (Supplementary Table S4¹⁷⁹) is similar to DH5α (Sigma Aldrich commercial information, see Supplementary Table S4¹⁷⁹). The strain is resistant to the antibiotic streptomycin.

We amplified plasmid libraries in SIG10-MAX[®], extracted their DNA, and transformed them into mutants derived from *E. coli* K-12 strain BW25113 that harbor chromosomal deletions of the *crp*, *fis*, or *ihfA* gene. We obtained these mutant strains from the KEIO collection¹⁷⁹ and used them for sort-seq experiments. The design, genetic parts, and assembly of the plasmid vectors we used in this study are available in the Supplementary material, as are all primers, TFBS sequences/libraries, strains, and plasmids.

Sort-Seq procedure

To explore the regulatory effects of each transcription factor on binding sites in the corresponding library, we constructed three plasmids, each of which enables the inducible expression of one of our three TFs. These are plasmids pCAW-Sort-Seq-V2-CRP, pCAW-Sort-Seq-V2-Fis, and pCAW-Sort-Seq-V2-IHF (**Supplementary Methods 3** and [Supplementary Table S1](#)). We cloned the TFBS libraries ([Supplementary Table S3](#)) into their respective plasmids and then transformed them into mutant strains lacking the corresponding TFs (Δcrp , Δfis , and Δihf , as listed in [Supplementary Table S2](#)). We induced TF expression using anhydrotetracycline (Atc) and, after overnight growth, performed cell sorting for cell populations. During sorting, we distributed cells into 13 equally-spaced logarithmic bins based on their fluorescence levels. We replicated each sort-seq experiment three times from three separate library transformations for each TF. To mitigate the impact of extrinsic noise (gene expression variation among cells ¹⁸⁰), we adhered to standard protocols by normalizing our GFP fluorescence measurements against mScarlet-I fluorescence values obtained from flow-cytometry assays ^{7,71,181,182}. We subsequently recovered the sorted cells from each bin in 50mL Falcon tubes containing 10 mL of LB medium supplemented with chloramphenicol and incubated them overnight at 37°C with shaking at 220 rpm. After this growth period, we re-sorted cell cultures from each bin to eliminate potential contaminants and ensure that the cell populations had preserved their fluorescence distributions. Following re-sorting, we extracted plasmids from the cell population of each bin. We amplified and barcoded the TFBS region from each population through a polymerase chain reaction (PCR). Lastly, we sequenced barcoded amplicons containing TFBS sequences, and used the sequencing results to calculate the regulation strengths of each TF to its TFBSs in the corresponding library. More details are provided in the Supplementary Material.

Regulation strengths

Due to gene expression and measurement noise, individual TFBS variants in a sort-seq experiment usually appear in more than a single bin, and their read count (frequency) varies among bins ^{71,72,121,153}. Following established practice ^{18,72,123}, we used a weighted average of these frequencies for each variant to represent the mean expression level caused by the variant. To facilitate the interpretation of this quantity, we converted this expression level into a regulation strength relative to the highest observed regulation strength for a given TF, to which we assigned a value of one. From each library, we selected a single naturally occurring binding site for each TF that was previously characterized in the literature as a strong binder. We called this TFBS the WT sequence and used it as a baseline to separate weakly (higher GFP expression) from strongly (lower GFP expression) regulating TFBSs.

Validating regulation strengths with plate reader measurements

To further validate our regulation strength data, we chose 10 DNA binding sites from each bin (10 variants $\times$ 13 bins = 130 variants in total per TF library, **Supplementary File 2**), covering a wide range of measured regulation strengths. We cloned these sequences into the appropriate vector pCAW-Sort-Seq-V2-TF (TF: CRP, Fis or IHF), and transformed them into the appropriate mutant strain. We picked individual colonies and grew them overnight (16 hours, 37°C, 220 rpm) in liquid LB supplemented with 50 μ g/mL of chloramphenicol and anhydrotetracycline. We diluted the cultures to 1:10 (v/v) in cold Dulbecco's PBS (Sigma-Aldrich #D8537) to a final volume of 1 mL. We transferred 200 μ L of the diluted cultures into individual wells in 96-well plates and measured GFP fluorescence (emission: 485nm/excitation: 510nm, bandpass: 20nm, gain: 50), as well as the optical density at 600nm (OD₆₀₀) as an indicator of cell density. We then normalized fluorescence by the measured OD₆₀₀ value to account for differences in cell density among cultures and compared the obtained ratios to the previously inferred regulation strengths for the 130 selected variants. We performed all such measurements in biological and technical triplicates (three colonies per sample, and three wells per colony, respectively).

Supplementary methods

1. General procedures

Although all general procedures have been previously described ¹⁸, we describe them again below for completeness.

1.1 Media and reagents

To prepare SOB medium, we mixed 25.5g of solid medium stock (VWR J906) with 960 ml of purified water and subjected the resulting suspension to autoclaving. To prepare the SOC medium, we dissolved 20 ml of 1 M D-glucose (Sigma G8270) and 20 ml of 1 M magnesium sulfate (Sigma 230391) in 960 ml of pre-prepared SOB solution. For the LB medium, we combined 25g of solid medium stock (Sigma-Aldrich L3522) with 1 liter of purified water and then autoclaved it. To prepare the M9 minimal medium, we diluted M9 minimal salt sourced from Sigma (M6030) in distilled water as per the manufacturer's guidelines, autoclaved the solution, and added 0.4% glucose (Sigma G8270), 0.2% casamino acid (Merk Millipore, 2240), 2 mM magnesium sulfate (Sigma 230391), and 0.1 mM calcium chloride (Sigma C7902). Where required, we supplemented growth media with chloramphenicol (50 µg/mL), anhydrotetracycline (100 ng/mL, Cayman-chemicals #10009542), and/or glucose (0.4% w/v final concentration). We prepared anhydrotetracycline by diluting the dried chemical in absolute ethanol, from a stock concentration (1000X) of 100 µg/mL to a working concentration of 100 ng/mL.

1.2 Overnight incubation of cultures in liquid and solid medium

We cultivated bacteria in liquid LB medium (using either 15mL or 50mL Falcon tubes), enriched with chloramphenicol at a concentration of 50 µg/mL. We incubated these cultures for a period of 16 hours at a temperature of 37°C, with a shaking speed of 200rpm and 50 mm orbital motion, using an Infors HT Multitron Incubator Shaker. Similarly, for cultures in solid medium, we grew bacterial colonies on LB-agar plates (using sterile plastic Petri dishes of 90mm × 15mm dimensions), also supplemented with chloramphenicol at a concentration of 50 µg/mL, and incubated them for the same time and at the same temperature.

1.3 PCR Reactions

Except where specifically mentioned, we amplified DNA fragments through a polymerase chain reactions (PCR), employing Q5[®] high-fidelity polymerase (NEB #M0491L) to minimize mutation introduction into the amplicons. We followed the protocol recommended by NEB, aiming for a final reaction mixture of 50µL. We conducted each PCR reaction twice, and combined the products from these duplicates after the completion of the reaction. We determined the primer melting temperatures (T_m) using the NEB T_m calculator (accessible at <https://tmcalculator.neb.com/#!/main> ) , with primers at a concentration of 500nM.

1.4 Verifying PCR products through gel electrophoresis

Unless indicated otherwise, we verified successful PCR amplification via gel electrophoresis to ensure the presence of singular-band amplicons and the absence of non-specific bands. This process involved separating PCR products in a 0.8% agarose Tris-EDTA (TAE) gel. We conducted electrophoresis for 45 minutes at 120V, or until the bands had progressed beyond the halfway point of the gel's total length.

1.5 DNA purification with commercial kits

Once we had confirmed a successful PCR through gel electrophoresis, we purified the PCR products utilizing the Monarch[®] DNA PCR/Gel Extraction Kit (NEB #T1020L), adhering to the manufacturer's protocol. In situations requiring further purification (such as the occurrence of non-specific bands post-PCR), we performed gel purification. This involved mixing 10 µL of 6x NEB DNA dye with each 50 µL PCR product, then loading all 60 µL onto a 1% agarose gel. We carried out electrophoresis for 45 minutes at 120V, or until the bands had moved more than halfway

through the gel. We then excised the DNA band corresponding to the amplified sequence using a scalpel. For extracting DNA from the gel, we used the Monarch[®] DNA Gel Extraction Kit (NEB #T1020L).

1.6 Gibson assembly¹⁸³

We assembled PCR-amplified fragments using the NEBuilder-HiFi[®] DNA Assembly Master Mix kit (NEB #E2621L). We determined the molarity required for assembly based on the guidelines outlined by the Barrick Lab (details available at <https://barricklab.org/twiki/bin/view/Lab/ProtocolsGibsonCloning>). We incubated the assembly mix for one hour at 50°C in a dry bath incubator, followed by cooling on ice for subsequent steps.

1.7 Preparation of electrocompetent cells

We utilized glycerol/mannitol step centrifugation to prepare electrocompetent cells¹⁸⁴. To this end, we first cultured the appropriate *E. coli* strain (Supplementary Table S4) in 5 mL SOB medium at 37°C with a shaking speed of 250 rpm overnight. The next day, we transferred 3 mL of the culture to 300 mL SOB medium, and incubated under the same conditions until the OD600 reached a value between 0.4 and 0.6 (measured at an optical path length of 1 cm), which took approximately 2-4 hours. After cooling the culture on ice for 15 minutes, we centrifuged the cells at 4°C and 1,500 g for 15 minutes. We then resuspended the cells in 60 mL of ice-cold distilled water (dH₂O) and divided them into three 50 mL tubes. Gradually, we added 10 mL of an ice-cold glycerol/mannitol solution (consisting of 20% glycerol (w/v) and 1.5% mannitol (w/v)) to each tube using a 10 mL pipette. We centrifuged the tubes at 1,500 g and 4°C for 15 minutes in an Eppendorf 5810/5810 R centrifuge with acceleration/deceleration set to zero. After discarding the supernatant, we resuspended cells in 3.0 mL of the same glycerol/mannitol solution. We transferred these suspensions to pre-cooled 1.5 mL tubes, and incubated them in a dry ice-ethanol bath for about 1 minute. Finally, we stored the suspensions at -80°C for future transformation experiments.

1.8 Electroporation

In all transformation experiments described in this study, we used 100 µL of electrocompetent cells for electroporation, employing 0.2 cm cuvettes (EP202, Cell Projects, UK) and a Micropulser electroporator (Bio-Rad) set to the EC3 setting (15k V/cm). Post-electroporation, we recovered cells in 1 mL of SOC media, warmed in advance in 15 mL Falcon tubes. This recovery step lasted for 1.5 hours at 37°C with a shaking speed of 220 rpm. Unless specified otherwise, we spread 300 µL of each recovered culture on an LB agar plate that contained 50 µg/mL chloramphenicol. We then incubated this plate overnight for 16 hours at 37 °C. Subsequently, we confirmed the identity of the clones on the plate via Sanger sequencing.

2. The design of plasmid pCAW-Sort-Seq-V2

The plasmid pCAW-Sort-Seq-V2 is a derivative of the plasmid pCAW-Sort-Seq¹⁸. Briefly, plasmid pCAW-Sort-Seq harbours a pBBR1 replication origin, which ensures a broad host range and a low copy number—typically between 5 to 10 copies per cell¹⁸⁵. It also harbours a chloramphenicol resistance gene, a TetR repression system¹⁸⁶ and a TFBS measuring module. The latter consists of an interchangeable TFBS that lies between a constitutive promoter and a superfolder GFP (*sfgfp*¹⁸⁷) reporter gene. In this system, TF-TFBS interactions can decrease GFP production by physically obstructing the bacterial RNA polymerase, a phenomenon known as steric hindrance. The reporter gene *sfgfp* is insulated by a transcriptional insulator named RiboJ, a synthetic ribozyme that removes 5'UTR interferences from variable TFBS sequences in the mRNA by self-cleavage¹⁸⁸. More details and features have been previously described in ref. ¹⁸. Here, we have integrated a bicistronic expression cassette into pCAW-Sort-Seq to enable the regulated expression of a TF of choice. This cassette harbors the gene encoding one of our focal TFs situated upstream of a *mscAlet-I*¹⁸⁹ reporter gene to monitor the expression of this bicistronic operon via fluorescence. Regulation is achieved through the *pLtetO-1*¹⁸⁶ promoter, a synthetic promoter tightly repressed

by TetR. Expression is initiated by the addition of anhydrotetracycline (Cayman Chemicals, catalog #10009542), which relieves TetR-mediated repression and activates the expression of the entire cassette.

3. Construction of the plasmid pCAW-Sort-Seq-V2 and its variants

We initiated the construction of pCAW-Sort-Seq-V2 by synthesizing (Twist Biosciences, California, USA) and cloning a codon-optimized *mscAlet-I* gene expression cassette into the pCAW-Sort-Seq plasmid. This gene is oriented antiparallel to the tetracycline resistance gene (*tetr*). This cloning step resulted in the formation of the intermediate pCAW-Sort-Seq-V2t plasmid (Supplementary Figure S1 [\[link\]](#)). Subsequently, we amplified for each of our three TFs the encoding gene (*crp*, *fis* and *ihf*) from the genome of the *E. coli* strain BW25113 and inserted it upstream of the *mscAlet-I* gene in a bicistronic operon configuration using Gibson assembly. This process produced TF-specific expression plasmid variants (Supplementary Figure S2 [\[link\]](#)). The final phase involved removing the core promoter to create negative controls for each TF and cloning libraries into each plasmid variant upstream of the *sfgfp* gene. This yielded the plasmids used in our experiments (Supplementary Table S1 [\[link\]](#), Supplementary Figure S2 [\[link\]](#)). The specific sets of primers used for each PCR reaction are listed in Supplementary Table S6 [\[link\]](#).

4. Library design, synthesis, and cloning

We based the design of the TFBS library for each TF on consensus sequences available in the RegulonDB database⁷⁰, the most comprehensive database for *E. coli* transcriptional studies. We designed each library by randomizing the eight most important TFBS positions, such that each of the four nucleotides had an equal probability (0.25) to occur at each position. The expected library size for this approach is $(4^8) = 65,536$ sequences. Library compositions can be found in Supplementary Table S3 [\[link\]](#). We designed libraries in the Snapgene[®] software ([snapgene.com](#) [\[link\]](#)), and had them synthesized by IDT (Coralville, USA) as single-stranded DNA Ultramers[®] of 140bp (4nmol). We resuspended each library in nuclease-free distilled water and serially diluted it to a concentration of 50ng/uL. We used Ultramers[®] as templates in a PCR reaction for the formation of dsDNA fragments and library amplification. We amplified the Ultramer[®] DNA molecules with the following PCR program: 98°C/30 s; 25 cycles of 98°C/10 s, 60°C/15 s and 72°C/80 s; and 1 cycle of 72°C/5 min. We opted for a maximum of 20 cycles to reduce amplification biases. We analyzed amplification products through gel electrophoresis to confirm that only a single product band with the expected size of around 140bp was present for each PCR reaction. After confirming the presence of single bands, we purified the products using the Monarch[®] DNA gel extraction kit (NEB #T1020L). Whenever unspecific bands appeared during electrophoresis, we gel-purified the PCR products using the same kit.

For the construction of each plasmid-based library, we digested 1μg of the purified library with HindIII-HF (NEB #R3104) and BamHI (NEB #R3136) restriction enzymes in a 100μL reaction, followed by overnight incubation at 37°C. We isolated the appropriate cloning plasmid from its host strain using the QIAprep spin miniprep kit (Qiagen, Germany), and digested it with the same enzymes. Post-digestion, we added 3μL of Quick CIP (calf intestinal alkaline phosphatase) to prevent self-ligation by dephosphorylating DNA ends. We purified the ligated DNA using the Monarch[®] DNA gel extraction kit (NEB #T1020L).

We performed the ligation with a 10:1 molar ratio of insert-to-vector, using 100ng of vector, 10 units of T4 DNA ligase (NEB #M0202L), and 2 μL of 10X ligation buffer in a 20 μL reaction. We incubated the mixture at 20-22°C for approximately 16 hours, followed by 15-minute inactivation of the ligase at 65°C. We purified the ligation product, resulting in 10 μL of DNA resuspended in dH₂O. We transformed *E. coli* SIG10-MAX[®] cells by electroporation with this purified product.

Post-transformation, we plated 50 μL of each recovered culture on LB agar for colony-forming unit counting (cfu) and transformation efficiency estimation. From the agar plates, we selected 30 colonies for colony PCR and Sanger sequencing to assess library diversity (NightSeq[®] service, Microsynth, Switzerland). Our mean transformation efficiency was 10⁶ cells per transformation.

We diluted the remaining 950 μL of transformants in 9 mL of LB medium with chloramphenicol and cultured it overnight. We then aliquoted the overnight culture into 1 mL cryotubes with 20% glycerol and stored at -80°C .

After transforming plasmid libraries into the SIG10-MAX® strains for reasons of transformation efficiency and plasmid maintenance, we extracted the plasmids using a QIAprep spin miniprep kit (Qiagen, Germany), and transformed them into the appropriate host mutant host strains Δcrp , Δfis and Δihf (see the strain genotypes in [Supplementary Table S2](#)). We cultured each library-transformed strain overnight, aliquoted it in 1 mL cryotubes with 20% glycerol, and stored it at -80°C for further experimentation.

5. Analysing and sorting cells

In preparation for cell sorting, we cultivated cells harboring each library in liquid LB medium enriched with chloramphenicol. Specifically, we cultured 1 mL of transformed cell aliquots and a streak of cells with a control plasmid (a promoterless pCAW-Sort-Seq-V2 plasmid lacking sfGFP expression) in 50 mL Falcon tubes with 9 mL LB medium (containing 50 $\mu\text{g}/\text{mL}$ chloramphenicol) overnight. Subsequently, we diluted these overnight cultures at a 1:100 ratio (v/v) in LB medium with chloramphenicol and divided them into two aliquots. We supplemented one aliquot with the anhydrotetracycline (Atc) inducer to express the plasmid-encoded TF. The other aliquot remained unchanged. We incubated both cultures for 5 hours until late-exponential/early-stationary phase (200 RPM, 37°C). Subsequently, we diluted 20 μL of the cultures in 1 mL of cold filtered Dulbecco's PBS (Sigma-Aldrich #D8537) in 15 mL FACS tubes.

We performed FACS-sorting on a FACS Aria III flow cytometer (BD Biosciences, San Jose, CA) using a 70 μm nozzle. We utilized a 488 nm laser for detecting forward scatter (FSC) and side scatter (SSC) with a 488 nm/10 nm band-pass filter. We set the flow rate to 1.0, adjusting sample dilution as necessary to achieve no more than $\approx 10,000$ events/second. Considering the small size of bacterial cells, we reduced the particle detection threshold to the lowest feasible setting (200 arbitrary units on FSC and SSC channels), increasing it to a maximum of 500 units if background noise was excessive. We then adjusted FSC-H and SSC-H for cells with the negative control plasmid to center the bacterial population in the cytometer's software (BD FACSDiva™ Software v9.0). We based the sorting and binning of cells on both mScarlet-I (PE-Texas-Red channel, excitation laser: YellowGreen 561 nm, LP filter: 600 nm, BP filter: 610/20 nm) and sfGFP fluorescence (FITC channel, excitation laser: 488 nm, LP filter: 502 nm, BP filter: 530 nm/30 nm), setting both the PE-Texas-Red and FITC channels voltages so that the median fluorescence of the negative control was between 0 and 100 (arbitrary units) on the FITC-H and PE-Texas-Red-H axes.

Initially, the sorting process involved establishing a gate for identifying cells that are red-fluorescence-positive, i.e., reporter and thus TF-expressing. To establish this gate, we began by measuring the autofluorescence of our negative control culture on the PE-Texas-Red-H axis. Thereafter, we examined the fluorescence of a positive control that had been induced with Atc and was expressing the mScarlet-I protein. We then configured a gate around the positive mScarlet-I-expressing population, ensuring that all cells sorted through this gate exhibited a level of reporter expression higher than the negative control.

Next, we established the green-fluorescent sorting gates. For setting sorting gates on the FITC-H axis, we first recorded the autofluorescence of the negative control culture. This median autofluorescence defined the upper boundary of the lowest bin (B1) for the experimental population. We then analyzed the fluorescence of 10^6 cells expressing sfGFP and containing the library, without sorting, in order to establish the boundaries of the next binning gates. We took the lower bound of the highest bin (B13) for the experimental population to correspond to the 95th percentile of the fluorescence distribution of this population. We chose boundaries between the remaining 11 intermediate bins with equidistant spacing on a binary logarithmic ($\log_2$) scale. After we had determined the gates in this way, we calculated the fractions of the previously 10^6 recorded cells that fell inside each bin.

To maintain statistical robustness and minimize sampling error in our downstream analyses, we wanted to ensure that each of the 65,536 unique sequences in our library was represented by at least 30 cells after sorting. In addition, it was crucial to account for an anticipated diversity loss of up to 70% of the cell population during sorting, due to factors such as cell death and the dilution of low-frequency and lower-fitness genotypes during the post-sorting recovery phase¹²². To establish the number of cells required to be sorted initially, we thus determined a multiplicative factor f based on the minimally needed number of 30 cells per sequence and the expected cell retention rate post-sorting of 30%, i.e., $f = \frac{30}{0.30} = 100$ cells/sequence. We thus multiplied our initial library size by 100 for sorting purposes, i.e., we sorted total of 6,553,600 cells. This calculation aims to ensure that despite a substantial reduction in cell numbers post-sorting, each unique sequence remains adequately represented in the surviving cell population. We sorted cells into 1.5 mL Eppendorf tubes, each containing 500 μ L of LB medium, and kept at 4 °C to prevent growth during sorting and sample processing. We replicated the entire sorting procedure three times, each based on independent library transformations.

After sorting, we added 1mL of LB without antibiotics to each tube and removed 20 μ L of the resulting volume for serial dilutions. We transferred the remaining liquid culture (980 μ L) to 50 mL falcon tubes and allowed each culture to recover for 2 hours (37°C, 220 rpm). After recovery, we added 9mL of LB supplemented with chloramphenicol, and grew the cultures overnight for freezing part of them in glycerol stock aliquots, for validating our binning procedure (see below), and for extracting plasmid DNA for subsequent PCR and sequencing steps. We used the 20 μ L initially removed from each 1mL culture for preparing two serial dilutions (10^{-4} and 10^{-6}) that we plated on LB-Cm agar plates (200 μ L per plate) to estimate the post-sorting viability through colony forming unit (cfu) counting. By knowing how many cells were sorted into each bin, we can estimate how many cells would be expected in our dilutions and compare this number with the cfu counts we observed. In this way, we estimated that on average (across bins) 77% of cells remained viable (standard deviation: 18.15%). We also estimated the genetic diversity of the library through Sanger sequencing of DNA from individual colonies (NightSeq® service, Microsynth, Switzerland).

To validate our binning procedure, we re-grew binned cultures from either overnight recovered cultures or frozen aliquot stocks. We then measured their expression distributions by flow cytometry, which reproduced the original expression measurements. This approach allowed us to compare the fluorescence distributions of the re-grown cultures with the pre-sorting distributions. Specifically, we assessed whether the geometric mean of the fluorescence distribution for each sorting bin matched those recorded during the initial sorting procedure. Next, we re-sorted cells derived from each bin in order to eliminate cell cross-contaminations and other factors that could alter the bin distributions. The geometric mean is often preferred over the arithmetic mean in this type of analysis, because it is less affected by the presence of outliers in the data^{190,191}. It is also a more accurate representation of the central tendency of data that is log-normally distributed, which is often the case with flow cytometry fluorescence measurements^{190,191}.

6. DNA extraction and sequencing

We diluted 500 μ L of individual glycerol stocks of each replicate subpopulation of sorted cells (i.e., cells from each bin of fluorescence intensity) in 5mL of LB supplemented with chloramphenicol in 15mL Falcon tubes, and grew the resulting cell culture overnight (16 hours, 37°C, 220 rpm). On the next day, we isolated plasmids from each culture using a QIAprep® spin miniprep kit (Qiagen, Germany). In order to allow the sequencing of multiple pooled samples (multiplexing), we barcoded our regulatory region through PCR with specific HPLC-purified primers (Supplementary Table S6¹⁹⁰) provided by Eurofins (Konstanz, Germany). We added barcodes to the 5' region of the amplicon through a PCR reaction. We performed this PCR reaction with the Q5 high-fidelity polymerase, and did so in triplicate for each miniprep-isolated plasmid library. To calculate primer melting temperatures (T_m), we used the NEB T_m calculator

(<https://tmcalsculator.neb.com/#!/main>) for a primer concentration of 500nM. We performed the PCR with the following program: 98°C/30 s; 25 cycles of 98°C/10 s, 64°C/30 s and 72°C/30 s; and 1 cycle of 72°C/2 min.

After PCR amplification, we digested the reaction products with the restriction enzymes DpnI (NEB #R0176L) and Exonuclease I (NEB #M0293L) in order to remove traces of genomic DNA, plasmids, and single-stranded DNA that could interfere with sequencing. The Master Mix we used for a single digestion harbored 1µL of 10x CutSmart® Buffer (NEB #B6004S), 1µL of Exonuclease I (NEB #M0293L), 1µL of DpnI restriction enzyme (NEB #R0176L), and 7µL of distilled nuclease-free water. For each PCR product, we added 10µL of the Master Mix. We incubated the reaction for 1 hour at 37°C, following 15 minutes at 80°C for deactivation of the enzymes.

We purified the digestion products using the Monarch® DNA PCR/Gel Extraction Kit (NEB #T1020L) and fractionated them through gel electrophoresis to confirm that only a single band with a size ≈150bp was present. After this confirmation, we pooled the purified PCR products from the different bins of each replicate sorting equimolarly to a total mass of 2,600 ng and a volume of 100 µL (26ng/µL of DNA) in 1.5mL Eppendorf tubes. We then sent the pooled purified amplicons for adapter ligation and sequencing at Eurofins (NGSelect Amplicons® on Illumina HiSeq), obtaining 15 million paired-end reads (2 × 150 bp) for all samples.

We quantified the total number of reads (t) required for sequencing based on several factors, including the number of bins (b), biological replicates (r), strains (s), genotypes (g), and the reads per genotype (p). Using these variables, we first calculated the total number n_{bc} of needed barcodes as follows:

$$n_{bc} = b \times r \times s$$

For each of our three TFs, the values of these parameters are $b=13$ bins, $r=3$ biological replicates, and $s=1$ strains, yielding $n_{bc}=39$. The total number of required sequence reads can be estimated through the equation:

$$t = r \times g \times p$$

where we aimed for $p=30$ paired-reads per genotype. The number g of expected genotypes per library is the same (65,536) for each TF. These values lead to a total of $t = 3 \times 65,536 \times 30 = 7,667,712$ required paired-end reads for each library.

7. Data analysis

7.1 Filtering and preparing sequencing reads

We processed the sequencing data with a blend of custom python and awk scripts, complemented by established bioinformatics utilities. Firstly, we trimmed sequences by computationally excising Illumina adapters, followed by merging paired-end reads. We then organized the paired-end reads into distinct files, each tagged with a sequence barcode identifying the sequence bin from which the corresponding reads originated.

For the removal of Illumina adapter sequences from the paired-end reads, we employed Cutadapt ¹⁹² with the following parameters:

```
cutadapt -j 8 -e 0.1 --no-indels --overlap=8 --discard-untrimmed \
-a "^{$FWD...}$REV_RC;max_error_rate=0.2;min_overlap=6" \
-A "^{$REV...}$FWD_RC;max_error_rate=0.2;min_overlap=6" --pair-filter=any \
-o 'Sample_${sample}_trimmed_1.fastq.gz' \

-p 'Sample_${sample}_trimmed_2.fastq.gz' \

--max-ee=2 -l=114 \

$reads
```

Here, “ADAPTER_FWD” and “ADAPTER_REV_RC” are placeholders for the actual adapter sequences used, which are 5'-AGATCGGAAGAGCACACGTCTGAACTCCAGTCA-3' for read 1, and 5'-AGATCGGAAGAGCGTCGTAGGGAAGAGTGT-3' for read 2.

Because our amplicons were short (153 bp) and each sequencing sample was divided into two files with overlapping single-end reads, merging the two files (single-end reads) was necessary, for which we used the FLASH software¹⁹³. After the filtering and merging of reads, we demultiplexed paired-end reads with the following FLASH parameters:

```
flash -t $ncores $reads -O -m 60 -M 140 -z -o 'Sample_${sample}_merged'
```

Next, we utilized the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)¹⁹⁴ to discern and preserve only those sequences that exceeded a high-quality threshold of Q = 33. Subsequently, we refined our dataset by removing sequences that contained undesirable mutations or insertions/deletions (indels) within the region extending from the start of the core promoter to the end of the variable TFBS library. We performed this filtering step using a custom R¹⁹⁵ script. Thereafter, we used custom awk and R¹⁹⁵ scripts to convert the data into a table. Each row of this table contains data from one TFBS in the library, and each column contains the number of reads for this TFBS from one of the thirteen bins into which we had sorted cells. We estimated the average regulation strength of each TFBS from this data.

7.2 Calculating regulation strengths

We calculated regulation strengths as previously described¹⁸. Briefly, in sort-seq experiments, sequences often appear in multiple fluorescence bins due to random mis-sorting^{71,117} and the stochastic nature of gene expression¹⁸⁰. Following methods established in previous research^{16,72,123}, we determined the reporter expression level driven from each TFBS variant by calculating a weighted average of bins in which the variant occurred. This involved multiplying the frequency of each sequence (x_i) in a given bin i by a numerical value representing that bin ($w_i=1,2,3,...,13$), and then averaging these products over the total sequence count. Mathematically, this weighted average calculates as

$$e = \frac{\sum_{i=1}^n (x_i * w_i)}{\sum_{i=1}^n x_i},$$

where e is the expression level driven by the sequence.

This approach yields a continuous spectrum of expression levels e within the range of 1 to 13. High expression values indicate robust GFP expression and, thus, weak binding of a TF to a TFBS variant. To facilitate interpretation, we inverted this scale so that higher values indicate stronger

repression and lower GFP expression, i.e., we define the regulation strength b (strength of regulation) as follows:

$$b = e_{max} + 1 - e$$

where $e_{max}=13$ is the maximal expression level (maximal bin value). We then normalized b by the maximal regulation strength among all TFBSs we studied for a given TF. The resulting normalized regulation strength scores b vary from 0 to 1, where low scores denote TFBS variants with weak TF binding (weak reporter repression), and high scores indicate variants with strong TF binding (strong reporter repression). A score of $b = 1$ signifies the highest regulation strength (repression) calculated among all TFBSs for a given TF.

7.3 Combining data from triplicates

As a quality-filtering step, we eliminated TFBS variants that did not appear in all three replicates or that were represented by less than 30 reads in total (summing their read counts over the thirteen bins). Given that we have 13 bins and most sequences are typically found within an average of 3.4 bins, we wanted a minimum of 10 reads per bin to prevent misinterpretation of repression levels and also guide our threshold's choice. This procedure reduced the fraction of TFBS variants for which we had sort-seq-based sequence data from 95%, 90%, and 93% of the total library size (65,536 sequences) for CRP, Fis and IHF, respectively, to 49%, 66% and 63%.

Subsequently, we calculated the regulation strengths b of the remaining TFBS variants (**Supplementary Materials and Methods 7.2**). Then, we determined the coefficient of variation of regulation strengths across replicates for each TFBS variant, which estimate the consistency of measured regulation strengths among replicates. Notably, most TFBS variants exhibited a coefficient of variation (CV) below 0.5, which was the threshold we set for filtering sequences. Sequences with a CV above 0.5 were excluded from the analysis. This threshold is commonly used in transcriptional studies employing fluorescent reporters to account for transcriptional noise and measurement variability¹⁸⁰. Finally, we averaged regulation strengths for each TFBS variant across all replicates and normalized the resulting averages by the maximal observed regulation strength among all TFBSs of a given TF.

7.4 Frequency matrices and sequence logos

We generated frequency matrices of nucleotides that occur in TFBSs from each fluorescence bin by counting the frequency of each nucleotide at each variable position of the TFBS library. From this data, we generated heatmaps and DNA sequence logos representing the frequency matrices graphically. A sequence logo consists of a stack of the letters A, C, G, and T, at each position of a DNA sequence, where the relative size of each letter indicates its frequency in the sequence. The total height of the stack corresponds to the information content of that position, in bits^{49,196}. A sequence logo is a graphical representation of the informational properties of a TFBS. When mutated, nucleotides with high information content are more likely to lead to a loss of binding (repression) than nucleotides with low information content.

7.5 Creation of genotype networks and determining network metrics

We used in-house Python script and the Python package *igraph*¹⁹⁷ to generate directed genotype networks. These are graphs in which TFBS variants are nodes (vertices, genotypes), and variants that differ in a single nucleotide are connected by an edge. Each node of this network is associated with the corresponding DNA sequence and the associated regulation strength. Each edge is directed, i.e., it corresponds to a binding-score-increasing mutation, and points from a TFBS variant with lower regulation strength to a neighbor with higher regulation strength. We extracted the largest weakly connected subgraph (the “giant component”⁶) of the network, and used it for all further analyses. This giant component comprises the vast majority (99%) of sequenced genotypes for all of our TF landscapes. We used in-house Python and R scripts for all network analyses described below.

Epistasis: Epistasis refers to non-additive interactions between two or more mutations. It can impose severe constraints on molecular evolution, because the mutations that are beneficial in one genetic background may be deleterious in another²⁴. Epistasis between two mutations can be classified as magnitude, simple sign, or reciprocal sign epistasis, depending on the sign (i.e., positive or negative) of the fitness effect of individual mutations and their combinations²⁴. In magnitude epistasis, the effect of a mutation on regulation strength varies depending on the genetic background but the sign of this effect (increasing or decreasing regulation strength) does not. Simple sign epistasis occurs if one single mutant has a lower regulation strength than both the wild type and the double mutant, while the other single mutant has a regulation strength that is intermediate to the wild type and double mutant. Reciprocal sign epistasis occurs when both mutations independently decrease regulation strength, but their combination increases regulation strength. The presence of reciprocal sign epistasis is a necessary condition for the existence of multiple peaks in an adaptive landscape^{93,94}.

To determine the incidence of epistasis in our landscapes, we employed a method that involves identifying all “squares” in a genotype network with the *motifs* function from the *igraph* library in R. Each square consists of a “wild-type” sequence, a double-nucleotide mutant, and the corresponding two single mutants. We assessed epistasis for each square along a single axis by selecting the highest- regulation strength sequence as the double mutant. Our analysis placed each square into one of three categories: no sign epistasis, simple sign epistasis, and reciprocal sign epistasis. The no sign epistasis category included both magnitude epistasis and additivity (no epistasis) without differentiating between them, because neither affects the accessibility of landscape peaks through only binding-increasing mutations^{11,108}. We determined the proportion of all squares that fell into each category (see [Supplementary Table S5](#) [↗](#)).

Peaks: A peak is a genotype (TF binding site variant) whose neighbors all convey lower regulation strength than itself. Two peaks are connected if they are neighbors and convey the same regulation strength. We refer to the genotype with the highest regulation strength as the summit or global peak^{6,198}.

Genotype connectivity: To evaluate the sparsity of our networks and the likelihood of peak misclassification due to the incompleteness of our landscapes, we used a metric called “genotype connectivity.” This represents the number of immediate (1-mutant) neighbors of a given genotype for which our experiment produced repression strength data. A perfect experiment creating a complete landscape data set would provide repression data for each of our 4^8 genotype and all of its 24 immediate neighbors. In practice, however, data for some genotypes and some of their neighbors is unavailable. We quantified the “relative connectivity” of any one genotype as the ratio of the actual number of adjacent genotypes for which we have repression data to the theoretical maximum of 24 neighbors. This measure helps us to address questions about the sparsity of our landscape and its implications. Specifically, it allows us to determine if our sample is biased, with some regions or genotypes being more connected than others. This is particularly important for assessing whether peaks and high peaks are less connected than non-peaks, which could potentially lead to peak misassignments.

Accessible Paths: In the context of our genotype networks, we call a mutational path accessible if the regulation strength increases with each mutational step^{6,198}. We systematically enumerated all shortest accessible paths, distinguishing them by their path length, i.e., by the number of mutational steps in a path. Notably, there can be multiple alternative accessible paths with the same number of mutational steps from a single starting genotype to a peak.

Basins of attraction: The basin of attraction of a peak comprises all TFBS variants from which accessible paths to the peak exist. We refer to the basin’s size as the number of variants in the basin. We determined basin sizes by exhaustive enumeration.

Overlap between basins. The basins of attraction of different peaks may comprise overlapping sets of variants. To determine the overlap between two basins B_1 and B_2 , we used the Jaccard index $J_{199,200}$, which is equal to the size of the intersection between two sets of variants divided by the size of their union:

$$J = \frac{B1 \cap B2}{B1 \cup B2}$$

7.6 Generating randomly shuffled landscapes

To generate uncorrelated random landscapes for our three TF landscapes, we performed a random shuffling of regulation strength values from our genotypes, followed by landscape analysis to identify peaks. To this end, we employed an ad-hoc python script to randomly permute experimentally measured regulation strength values among all genotypes in each of our landscape. This procedure preserves the distribution of regulation strengths in each landscape, but also creates an uncorrelated random landscape^{91,97}. Following random shuffling, we analyzed each shuffled landscape's topography with the same methods we had applied to the original (unshuffled) landscape to identify its peaks. To ensure statistical robustness, we repeated the random shuffling and peak identification 10^3 times. We then compared the number of peaks between the 10^3 shuffled landscapes and the original (unshuffled) landscape.

7.7 Principal component analysis

In order to investigate which nucleotides in a TFBS are most important for changes in regulation strengths, we performed a principal component analysis for each landscape (Supplementary Figures S19–S21). To this end, we first one-hot-encoded our data. This encoding represents each categorical value (nucleotide at each position of a DNA string) as a binary vector of length 4. It thus converts an entire DNA string of length L into a $4 \times L$ binary matrix. We used this binary matrix to perform PCA with the R base function *prcomp*.

8. Simulated adaptive walks

We simulated the adaptive evolution of a population on our landscapes by performing two different types of random walks, “greedy” adaptive random walks, and random walks based on Kimura's model of fixation probabilities.

For all simulations, we assumed that only point mutations occur and that the time it takes for a point mutation to become fixed in a population is much shorter than the time it takes for a new mutation to appear that will eventually go to fixation. This scenario is also called the strong selection weak mutation (SSWM) scenario^{16,99–101}. Under this scenario, evolving populations are monomorphic most of the time, i.e., all individuals have the same genotype. This scenario is realistic when the product of effective population size N and mutation rate μ is small ($N\mu < 1$), which is the case for *E. coli* ($N=1.8 \times 10^8$, $\mu=2 \times 10^{-10}$)⁶⁷. It allows us to model adaptive evolution as an adaptive random walk in our landscapes. Our simulations also account for mutation bias, which means that different types of mutation have a different probability of occurring in a population. We use mutation biases that were experimentally determined for *Escherichia coli*¹⁰⁴.

In a greedy adaptive random walk, starting from any one genotype, only the mutational neighbor that conveys the largest fitness advantage is fixed in a population. We initiated one greedy random walk from each non-peak genotype and terminated the walk once a fitness peak was reached, i.e., once every neighbor of the current genotype had lower fitness than the genotype itself. Because every greedy random walk is deterministic in the absence of neutral mutations, it is sufficient to initiate one greedy random walk per starting genotype.

Another well-established model for random walks allows for genetic drift. It uses fixation probabilities computed by Kimura^{98,201,202} i.e., $f_{ij} = (1 - e^{-2s}) / (1 - e^{-2Ns})$, where f_{ij} is the probability of fixing mutation j in the background of genotype i , N is the effective population size, and s is the selection coefficient, i.e. the difference in fitness between genotypes i and j ^{98,201}. For a given pair of genotypes, the only parameter of this model is the effective population size N , for which we explored values of $N=10^8$, $N=10^5$ and $N=10^2$. Generally, the smaller the population size is, the larger is the probability that neutral or deleterious mutations become fixed in a population.

For these “Kimura” adaptive walks, we chose 15,000 random starting genotypes for each of the three values of N , and simulated 1,000 random walks for each of them. At each step, we randomly picked a mutation j , generated a random number in the interval $[0, 1]$, and considered the

mutation to become fixed if the random number fell within the interval $[0, f_{ij}]$. Computationally, we accelerated this process by precomputing all fixation probabilities for all genetic backgrounds, and then used the *python* function *numpy.choice* to generate a random sample from the multinomial distribution of fixation probabilities at each step²⁰³. Because of genetic drift, Kimura's random walks do not necessarily terminate when they reach a fitness peak. For this reason, we simulated each such random walk for a maximum of 25 mutational steps.

9. Randomized landscape null model for peak accessibility

To evaluate whether the accessibility of strong regulatory peaks observed in the empirical landscapes exceeds random expectations, we constructed randomized “null” landscapes and repeated evolutionary simulations on them under identical conditions as for the empirical landscapes.

Specifically, for each transcription factor landscape, we generated 10^5 independently shuffled landscapes (see **Supplementary Methods 7.6**) by permuting repression-strength values across genotypes while preserving the network of genotypes and its mutational neighborhood structure. This procedure maintains the number of genotypes, their connectivity, and the distribution of repression-strength values, while removing correlations between genotype and phenotype. All other aspects of the analysis were kept identical to those used for the empirical landscapes.

For each randomized landscape, we simulated adaptive evolution using the same Kimura random-walk framework as for the empirical data. We assumed a population size of 10^8 individuals, initiated adaptive walks from the same number of starting genotypes, and used identical stopping criteria, fixation probabilities, and mutational bias parameters as described in **Supplementary Methods 8** (“*Simulated adaptive walks*”).

For each randomized landscape, we computed the fraction of adaptive walks that reached a high regulatory peak. This yielded a null distribution of peak-accessibility values for each transcription factor. We tested the null hypothesis that the observed fraction from the corresponding empirical landscape is greater than the mean fraction derived from this null distribution using a one-sided Monte Carlo permutation test ($N = 10^5$ randomized landscapes). For this test, we used Z-scores computed as the difference between the observed accessibility and the mean of the shuffled-landscape distribution, normalized by the standard deviation of that distribution.

10. Incorporating experimental uncertainty into adaptive walks

Fluorescence-based sort-seq measurements are subject to experimental uncertainty arising from finite fluorescence bin resolution, variability in fluorescence measurements, and sequencing noise. To account for this uncertainty when defining fitness relationships between genotypes and identifying peaks in the landscape, we incorporated empirically estimated noise into pairwise repression strength comparisons.

For each genotype, we computed repression strength (S) as a continuous quantity derived from the distribution of sequencing reads across fluorescence bins (see Methods). We quantified experimental uncertainty in S by a genotype-specific noise parameter (τ_G), calculated as the standard deviation of repression strength across three biological replicates. This parameter defines an uncertainty interval within which repression strength values cannot be reliably distinguished.

We compared the repression strengths of pairs of genotypes while explicitly accounting for their associated experimental uncertainty. Let S_A and S_B denote the repression strengths of genotypes A and B , with corresponding uncertainty estimates τ_A and τ_B . We considered genotype A to exhibit significantly stronger repression than genotype B if

$$S_A - \tau_A > S_B + \tau_B.$$

In this case, we assigned a directed edge from A to B in the genotype network. Conversely, genotype A was considered to exhibit significantly weaker repression than genotype B if

$$S_B - \tau_B > S_A + \tau_A,$$

in which case we assigned a directed edge from B to A . If neither condition was satisfied—i.e., if the uncertainty intervals of the two genotypes overlapped—we considered their repression strengths to be indistinguishable within experimental uncertainty. Such genotype pairs were treated as neutral and connected by two directed edges, one in each direction.

Applying these rules to all genotype pairs yielded the uncertainty-aware genotype network, denoted $\mathcal{G}^{S,\tau}$. For comparison, we also constructed a noise-free genotype network, denoted $\mathcal{G}^S$, in which experimental uncertainty was ignored ($\tau_G = 0$) and pairwise comparisons were based solely on point estimates of repression strength S . In both networks, peaks were defined as genotypes without outgoing edges.

Supplementary tables

Name	Antibiotic to select for plasmid persistence (concentration in $\mu\text{g/ml}$)	Relevant features	Source
pCAW-Sort-Seq-V2	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , no TFBS, no TF	This study, derived from ¹⁸
pCAW-Sort-Seq-V2-CRP-CRP _{WT}	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , CRP _{WT} operator, <i>crp</i>	This study
pCAW-Sort-Seq-V2-CRP-CRP _{lib}	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , CRP TFBS library, <i>crp</i>	This study
pCAW-Sort-Seq-V2-CRP-Neg	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , promoterless <i>sfgp</i> , <i>crp</i>	This study
pCAW-Sort-Seq-V2-Fis-Fis _{WT}	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , Fis _{WT} operator, <i>fis</i>	This study
pCAW-Sort-Seq-V2-Fis-Fis _{lib}	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , promoterless <i>sfgp</i> , <i>fis</i>	This study
pCAW-Sort-Seq-V2-Fis--Neg	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , Fis TFBS library, <i>fis</i>	This study
pCAW-Sort-Seq-V2-IHF-IHF _{WT}	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , IHF _{WT} operator, <i>ihf</i>	This study
pCAW-Sort-Seq-V2-IHF-IHF _{lib}	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , IHF TFBS library, <i>ihf</i>	This study
pCAW-Sort-Seq-V2-IHF-Neg	Chloramphenicol (50)	pBBR1, TetR, <i>sfgfp</i> , <i>mscarlet-I</i> , promoterless <i>sfgp</i> , <i>ihf</i>	This study

Table S1. Plasmids used in this study

Table S2. Table of strains

Strain	Genotype	Antibiotic resistance	Reference
SIG10-MAX from Sigma Aldrich <i>Cloning strain</i>	F- mcrA Δ (mrr-hsdRMS-mcrBC) endA1 recA1 Φ 80dlacZ Δ M15 Δ lacX74 araD139 Δ (ara,leu)7697galU galK rpsL nupG λ - tonA (StrR)	Streptomycin	Sigma Aldrich
<i>E. coli</i> DH5 α <i>Cloning strain</i>	F- endA1 gln V44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ 80dlacZ Δ M15 Δ (lacZYA- argF)U169, hsdR17(rK – mK +), λ -	None	204
<i>E. coli</i> JW5702-4 <i>Δcrp mutant strain.</i>	Δ (araD-araB)567, Δ lacZ4787(::rrnB-3), λ -, Δ crp- 765::kan, rph-1, Δ (rhaD- rhaB)568, hsdR514.	Kanamycin	179
<i>E. coli</i> JW1702-1 <i>ΔihfA mutant strain.</i>	Δ (araD-araB)567, Δ lacZ4787(::rrnB-3), λ -, Δ ihfA786::kan, rph-1, Δ (rhaD- rhaB)568, hsdR514.	Kanamycin	179
<i>E. coli</i> JW3229-1 <i>Δfis mutant strain</i>	b Δ (araD-araB)567, Δ lacZ4787(::rrnB-3), λ -, Δ fis- 779::kan, rph-1, Δ (rhaD- rhaB)568, hsdR514.	Kanamycin	179

Supplementary Table S3. Libraries used in the present study.

Sequence Name	Sequence*	TF	Library size
CRP	TGTNNTCTAGANNNTNTT	CRP-Fis	4 ⁸ = 65,536
Fis	NCTCNTTNNNAAGNNA	CRP-IHF	4 ⁸ = 65,536
IHF	TNNNNNATGAGNNN	Fis-IHF	4 ⁸ = 65,536

*IUPAC nucleotide nomenclature for degenerate nucleotides: N: A, C, T or G

Supplementary Table S4. TFBS reference (“wild-type”, WT) sequences used in the present study.

Sequence Name	Sequence	TF	Reference
WT _{CRP}	TGTGATCTAGATCACATTTT	CRP	85,205
WT _{Fis}	GCTCATTTTTTAAGCAA	Fis	83,206,207
WT _{IHF}	TAGTTTATGAGGTA	IHF	84,87

	<i>Number of genotypes¹</i>	<i>Number of peaks²</i>	<i>Number of low peak³</i>	<i>Number of high peaks⁴</i>	<i>Number of squares⁵</i>	<i>Magnitude epistasis or additivity^{6,7}</i>	<i>Simple sign epistasis⁷</i>	<i>Reciprocal sign epistasis⁷</i>
CRP	31,975 (49%) ¹	2,154 (7%)	2,093 (97.2%)	61 (2.8%)	580,289	37%	34%	28%
Fis	43,222 (66%) ¹	2,312 (5%)	2,140 (92.6%)	172 (7.4%)	1,194,144	34%	34%	32%
IHF	41,325 (63%) ¹	2,453 (6%)	2,254 (91.9%)	199 (8.1%)	1,026,818	34%	33%	32%

¹ Percentages indicate the proportion of the studied library relative to its expected size.

² Percentages indicate the proportion of peaks relative to the total number of genotypes

³ Low peaks are peaks with regulation strengths below the wild sequence. Percentages refer to the proportion of all peak genotypes.

⁴ High peaks are peaks with regulation strengths above the wild-type sequence. Percentages refer to the proportion of all peak genotypes.

⁵ A square represents a quadruplet of sequences containing a focal sequence (ab), one double mutant of this sequence (AB), and the two intermediate single mutants (Ab and aB).

⁶ This category includes both magnitude epistasis and additivity (no epistasis) without distinguishing them, because neither of the two subcategories affects peak accessibility^{11,108}

⁷ Percentages refer to the proportion of all squares.

Table S5. Global landscape properties

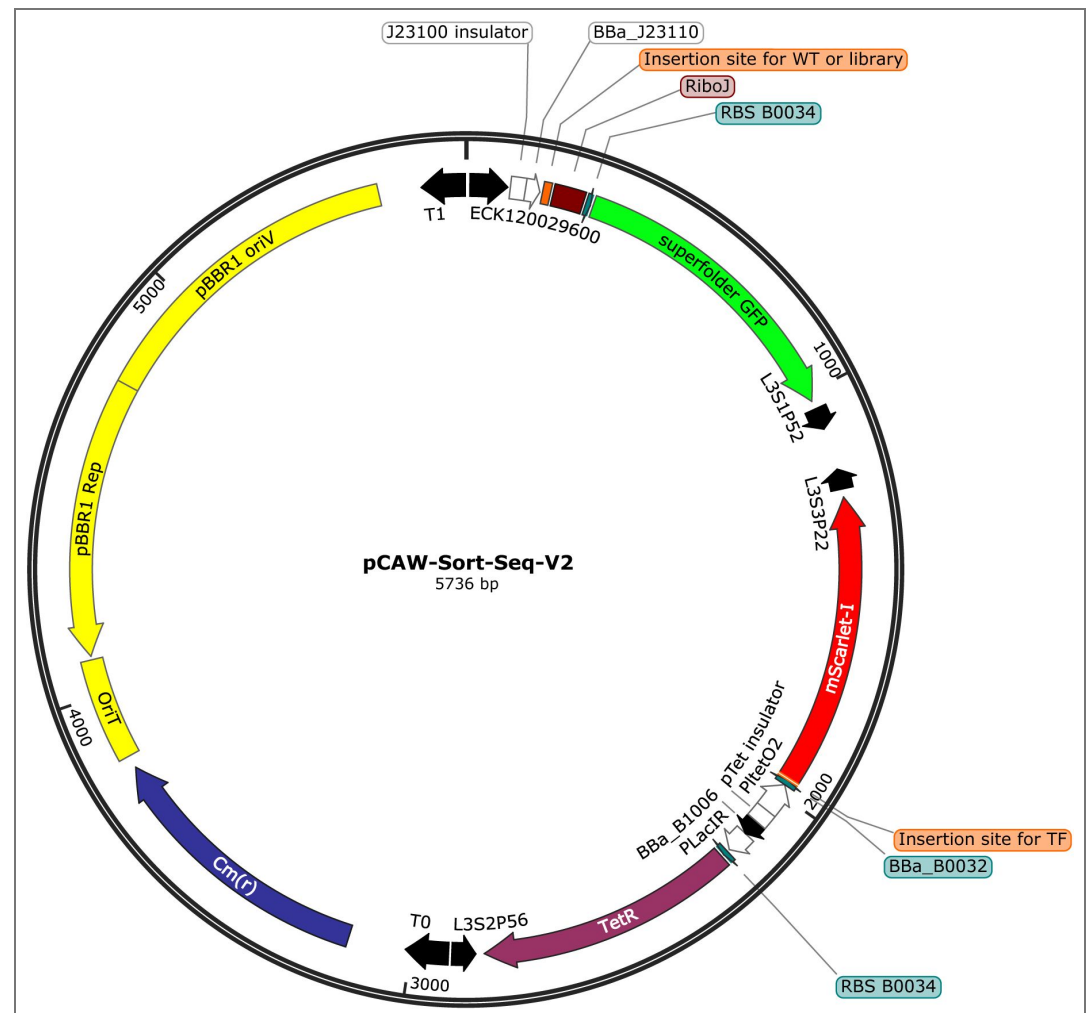
Name	Sequence	Usage
pCAW_frag1_F	5'- CGTCCGACTTACGGAAG GTAGATTTTACGGC-3'	Linearizing the pCAW-Sort-Seq fragment1 for Gibson Assembly
pCAW_frag1_R	5'- CTCGTGCCTAACGGAAG GTAGATTTTACGGC-3'	Linearizing the pCAW-Sort-Seq fragment1 for Gibson Assembly
pCAW_frag2_F	5'- TAAGATTGCCACGGAAG GTAGATTTTACGGC-3'	Linearizing the pCAW-Sort-Seq fragment2 for Gibson Assembly
pCAW_frag2_R	5'- AGGCCTGACTACGGAAG GTAGATTTTACGGC-3'	Linearizing the pCAW-Sort-Seq fragment2 for Gibson Assembly
Ultramer_ds_F	5'- TTCTCAAAGCTTCCTG CAGTATTC-3'	Amplifying Ultramer® libraries
Ultramer_ds_R	5'- CGGAAAGCACATCCGGT GAC-3'	Amplifying Ultramer® libraries
TFBS_R	5'- CCGTTTGTAGCATCACC TTC-3'	Sequencing the TFBS region
pCAW_Gibs_Lib_F	5'- GTCTGATGAGTCCGTGA GGACG-3'	Linearizing the pCAW-Sort-Seq plasmid
pCAW_Gibs_Lib_R	5'- GAGAAAAGAAAACCGC CGATCCTG-3'	Linearizing the pCAW-Sort-Seq plasmid
Ultramer_Gibs_F	5'- GGTGGACAGGATCGGCG GTTTTCTTTTCTTCTC AAAAGCTTCCTGCAGTA TTC-3'	Amplifying Ultramer® libraries for Gibson Assembly

Table S6. Primers for engineering the pCAW-Sort-Seq-V2 plasmid and for cloning libraries.

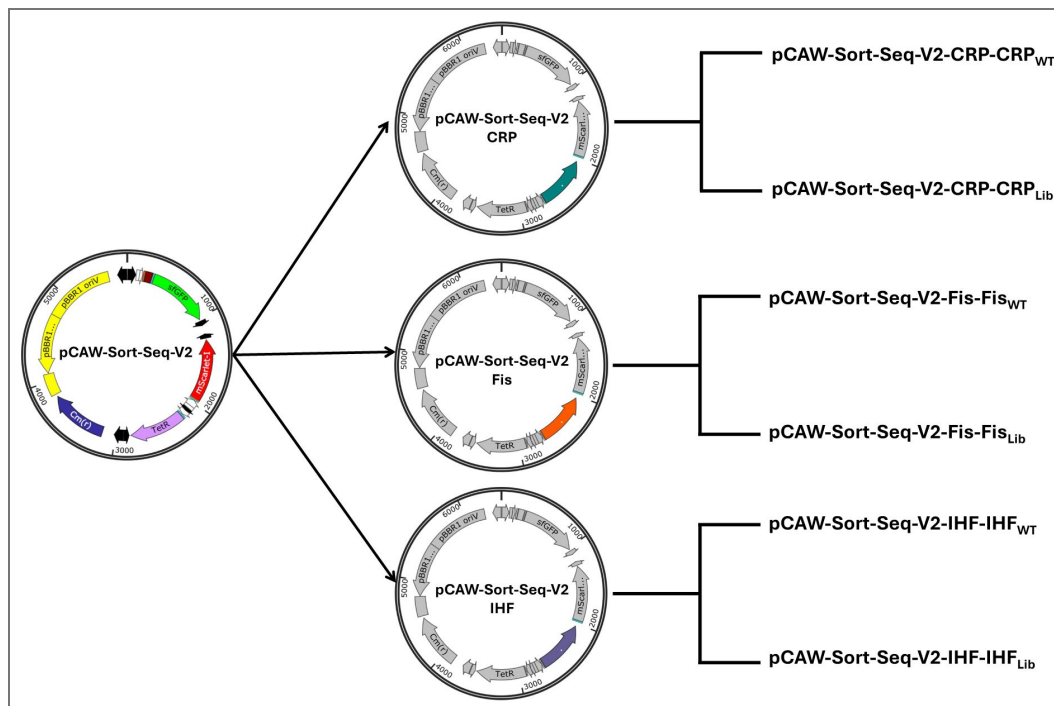
Ultramer_Gibs_R	5'- GGCTGTTTCGTCCTCAC GGACTCATCAGACCGGA AAGCACATCCGGTG-3'	Amplifying Ultramer® libraries for Gibson Assembly
-----------------	--	--

Table S6. (continued)

Supplementary figures

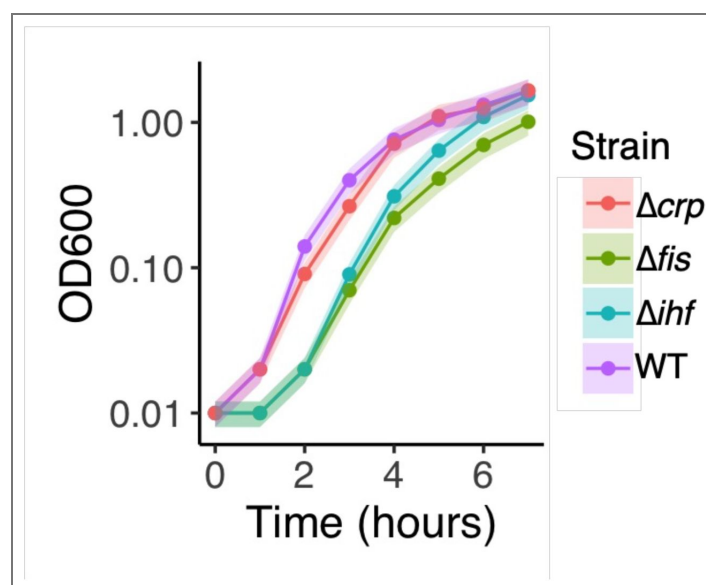


Supplementary Figure S1. Components of plasmid pCAW-Sort-Seq-V2. The plasmid pCAW-Sort-Seq-V2 is a derivative of the plasmid pCAW-Sort-Seq¹⁸. It encodes the broad-host-range, low-copy-number replication origin pBBR1¹⁸⁵ (in yellow, approximately 5–10 copies per cell¹⁸⁵), together with an origin of transfer (OriT) for conjugation (yellow), and a chloramphenicol resistance gene (dark blue). The interchangeable regulatory region where the TF binding site is inserted (orange) is positioned between a constitutive promoter (BBa_J23110²⁰⁸) and a superfolder GFP (*sfGFP*, green) fluorescent reporter gene¹⁸⁷. The transcriptional insulator RiboJ¹⁸⁸ (dark red) is located upstream of the *sfGFP* gene. The *tetR* gene (purple), derived from the original Tn10 transposon¹⁸⁶, is expressed under the control of a low-strength constitutive promoter (a pLac promoter variant²⁰⁹). TetR regulates the expression of a bicistronic operon encoding the TF of interest and the red fluorescent reporter *mScarlet-I*¹⁸⁹. The TF coding region is inserted at the TF insertion site (orange). Promoters and their associated insulators²¹⁰ are represented as white arrows, transcriptional terminators as black arrows²¹¹, and ribosome binding sites (RBSs) as small teal rectangles.



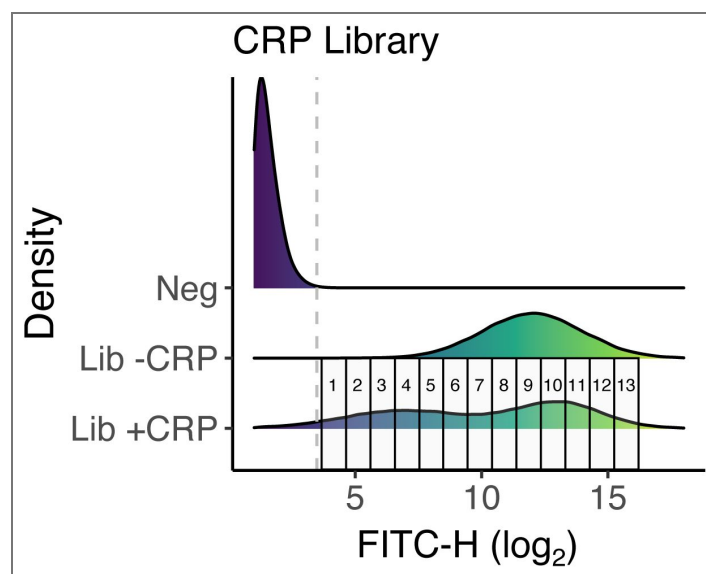
Supplementary Figure S2. Construction of pCAW-Sort-Seq-V2 global TF derivatives.

The plasmid pCAW-Sort-Seq-V2¹⁸ was derived from pCAW-Sort-Seq through a series of sequential cloning steps (see Supplementary Methods). First, a codon-optimized *mscAlet* expression cassette was synthesized and cloned into the parental pCAW-Sort-Seq backbone in an orientation antiparallel to the tetracycline resistance gene (*tetR*), generating the intermediate plasmid pCAW-Sort-Seq-*mscAlet*. Next, individual global transcription factor genes were inserted upstream of *mscAlet* to form a bicistronic operon, yielding three TF-specific intermediate plasmids expressing CRP, Fis, or IHF (CRP shown in green, Fis in orange, and IHF in purple; all other plasmid features are shown in grey for clarity). In the final step, depicted here, the appropriate wild-type TF binding sites and corresponding TFBS libraries were cloned upstream of the *gfp* reporter gene in each TF-specific backbone, producing the final library plasmids used in the sort-seq experiments. An additional negative control without a promoter region for the *sfGFP* gene was created for each TF (not shown). For full cloning details, see Supplementary Methods 1-4.



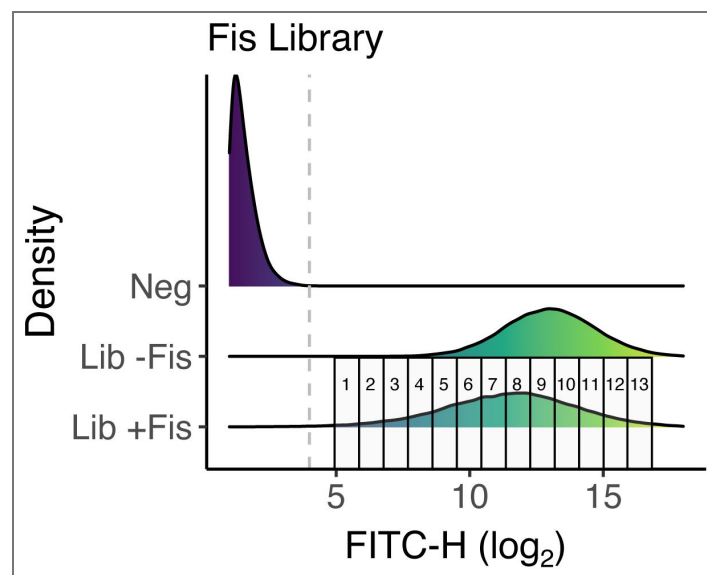
Supplementary Figure S3. Growth profile for strains used in this study.

Growth profiles for all strains used for landscape mapping are represented as optical densities (OD₆₀₀, y-axis) measured over 7 hours of growth at 1-hour intervals (x-axis). We transformed all strains with the plasmid pCAW-Sort-Seq-V2 and grew them in LB medium without inducers. Colors (legend) represent the different strains. Wild-type (purple) refers to the BW25113 parent-strain from which the mutants (Δcrp , Δfis , Δihf) were derived. Each circle represents the mean of three biological replicates independently measured in a plate reader. Shaded areas correspond to one standard deviation calculated from the three replicates.



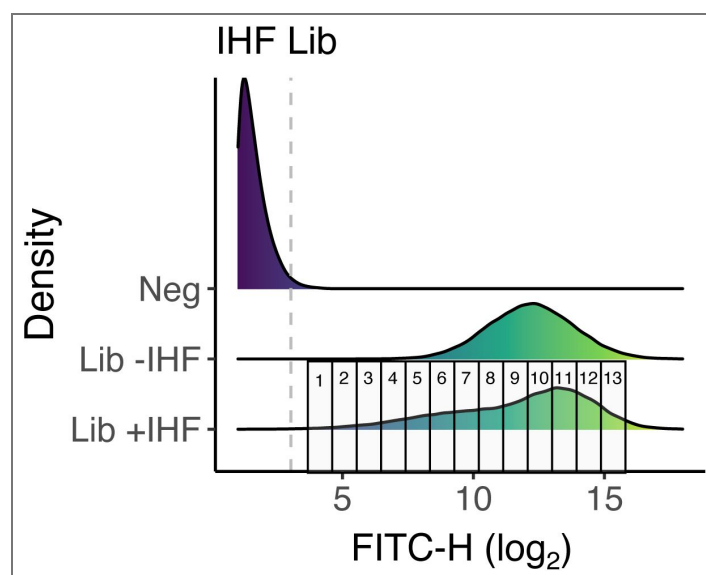
Supplementary Figure S4. Sorting procedure for CRP library.

Each density plot is based on a histogram representing the distribution of GFP fluorescence for 10^6 cells. The histogram represents data from one replicate, and the same sorting strategy was applied to the other two replicates. Specifically, the horizontal axis shows GFP fluorescence (FITC-H values, arbitrary units, $\log_2$ scale), while the vertical axis indicates the relative frequency of cells with a given fluorescence value. Heatmap colors indicate fluorescence, from low (dark purple) to high (yellow). The vertical axis indicates the three experimental conditions under which we measured fluorescence: 1) negative control ("Neg"), without GFP expression (promoterless pCAW-v2); 2) CRP library in pCAW-v2 without TF expression ("Lib -CRP") and 3) CRP library in pCAW-v2 with TF expression ("Lib +CRP"). To create each density plot from the histogram data, we performed density smoothing using a Gaussian kernel function with the aid of the *ggplot2* package^{21,22}. The vertical dashed grey line marks the upper limit of cell autofluorescence (negative control). We did not select cells with fluorescence below this threshold for sorting. Rectangular boxes (1 to 13) represent fluorescence bins/gates into which we sorted *E.coli* cells, spaced equally on a binary logarithmic ($\log_2$) scale.



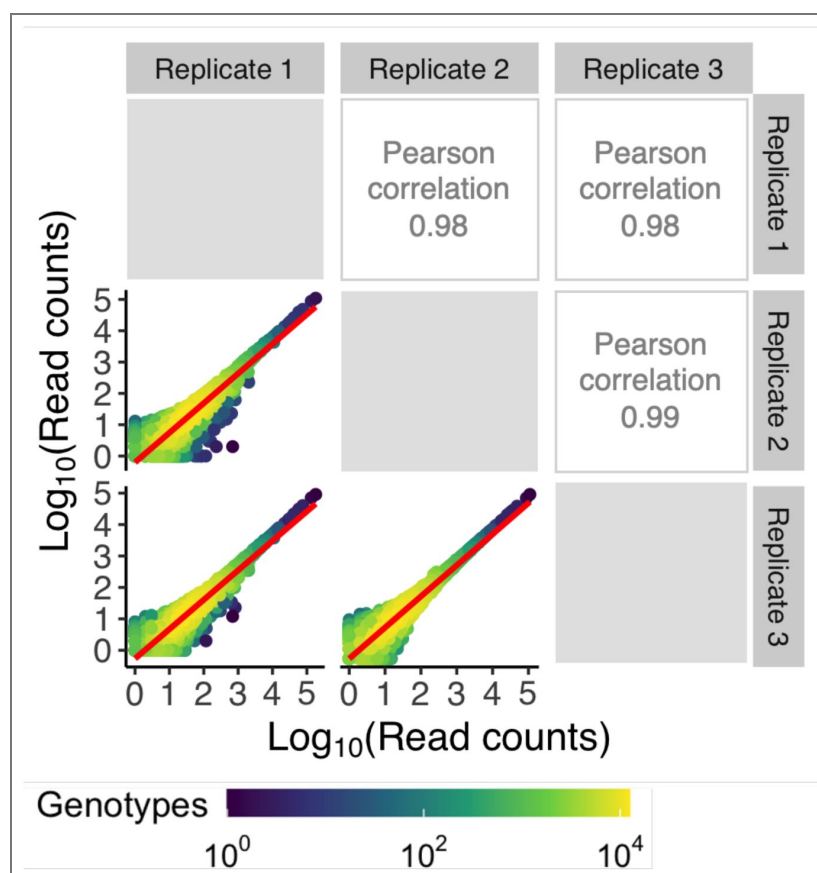
Supplementary Figure S5. Sorting procedure for Fis library.

Each density plot represents the distribution of GFP fluorescence for 10^6 cells. The histogram represents data from one replicate, and the same sorting strategy was applied to the other two replicates. Specifically, the horizontal axis shows GFP fluorescence (FITC-H values, arbitrary units, $\log_2$ scale), while each histogram indicates the relative frequency of cells with a given fluorescence value. Heatmap colors indicate fluorescence, from low (dark purple) to high (yellow). The vertical axis indicates the three experimental conditions under which we measured fluorescence: 1) negative control ("Neg"), without GFP expression (promoterless pCAW-v2); 2) Fis library in pCAW-v2 without TF expression ("Lib -Fis") and 3) Fis library in pCAW-v2 with TF expression ("Lib +Fis"). We performed density smoothing using a Gaussian kernel function to create a smooth density plot for each histogram with the aid of the *ggplot2* package^{21,22}. The vertical dashed grey line marks the upper limit of cell autofluorescence (negative control). We did not select cells with fluorescence below this threshold for sorting. Rectangular boxes (1 to 13) represent bins/gates for each sample, spaced equally on a binary logarithmic ($\log_2$) scale.



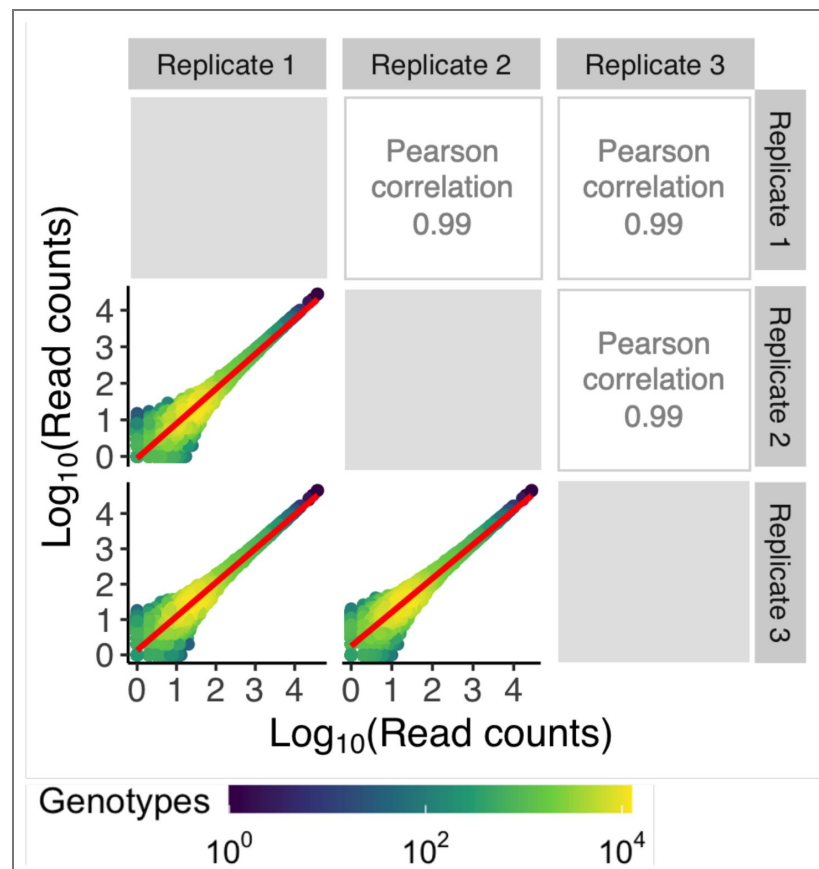
Supplementary Figure S6. Sorting procedure for IHF library.

Each density plot represents the distribution of GFP fluorescence for 10^6 cells. The histogram represents data from one replicate, and the same sorting strategy was applied to the other two replicates. Specifically, the horizontal axis shows GFP fluorescence (FITC-H values, arbitrary units, $\log_2$ scale), while each histogram indicates the relative frequency of cells with a given fluorescence value. Heatmap colors indicate fluorescence, from low (dark purple) to high (yellow). The vertical axis indicates the three experimental conditions under which we measured fluorescence: 1) negative control ("Neg"), without GFP expression (promoterless pCAW-v2); 2) IHF library in pCAW-v2 without TF expression ("Lib -IHF") and 3) IHF library in pCAW-v2 with TF expression ("Lib +IHF"). We performed density smoothing using a Gaussian kernel function to create a smooth density plot for each histogram with the aid of the *ggplot2* package^{21,12}. The vertical dashed grey line marks the upper limit of cell autofluorescence (negative control). We did not select cells with fluorescence below this threshold for sorting. Rectangular boxes (1 to 13) represent bins/gates for each sample, spaced equally on a binary logarithmic ($\log_2$) scale.



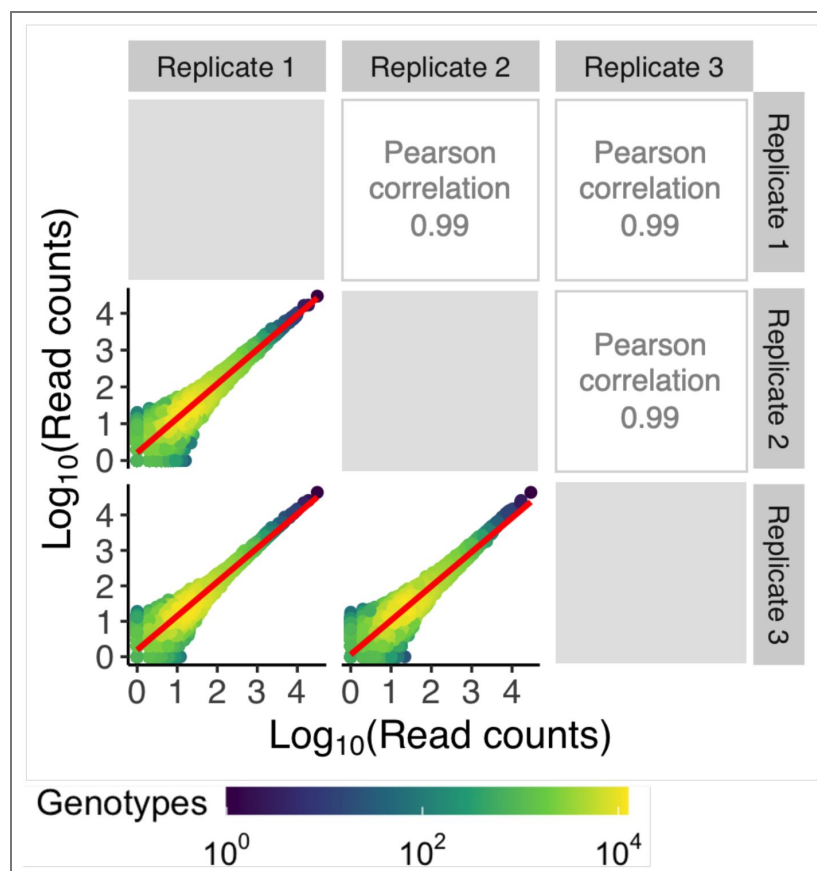
Supplementary Figure S7. Correlation of read counts among replicates of the CRP library.

The figure presents a pairwise comparison of log-transformed read counts across three experimental replicates for CRP. Each circle represents a distinct genotype, plotted according to its read counts in two sequencing replicates (x-axis for one replicate and y-axis for the other). The density of genotypes within the plot is indicated in the legend, with purple representing lower-density regions and yellow representing higher-density regions. The red line across each plot depicts a linear model's fit to the data, illustrating the association between the read counts of genotypes across replicates. Pearson correlation coefficients (upper triangle) are indicated for each pair of replicates. Their values are close to 1, demonstrating high reproducibility.



Supplementary Figure S8. Correlation of read counts among replicates of the Fis library.

The figure presents a pairwise comparison of log-transformed read counts across three experimental replicates for Fis. Each circle represents a distinct genotype, plotted according to its read counts in two sequencing replicates (x-axis for one replicate and y-axis for the other). The density of genotypes within the plot is indicated in the legend, with purple representing lower-density regions and yellow representing higher-density regions. The red line across each plot depicts a linear model's fit to the data, illustrating the association between the read counts of genotypes across replicates. Pearson correlation coefficients (upper triangle) are indicated for each pair of replicates. Their values are close to 1, demonstrating high reproducibility.

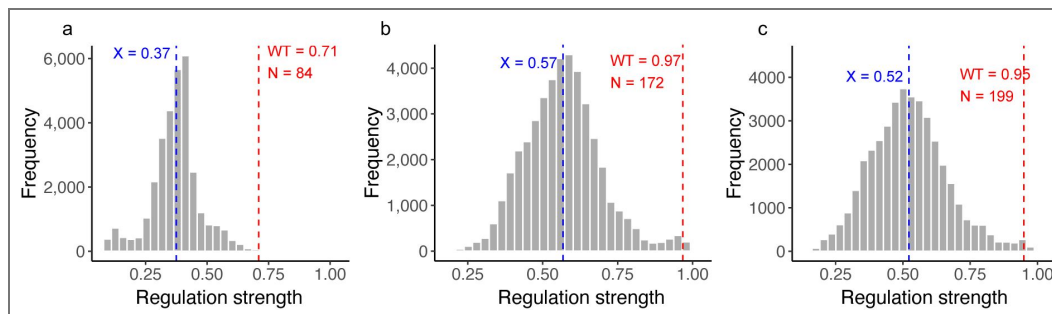


Supplementary Figure S9. Correlation of read counts among replicates of the IHF library.

The figure presents a pairwise comparison of log-transformed read counts across three experimental replicates for IHF. Each circle represents a distinct genotype, plotted according to its read counts in two sequencing replicates (x-axis for one replicate and y-axis for the other). The density of genotypes within the plot is indicated in the legend, with purple representing lower-density regions and yellow representing higher-density regions. The red line across each plot depicts a linear model's fit to the data, illustrating the association between the read counts of genotypes across replicates. Pearson correlation coefficients (upper triangle) are indicated for each pair of replicates. Their values are close to 1, demonstrating high reproducibility.

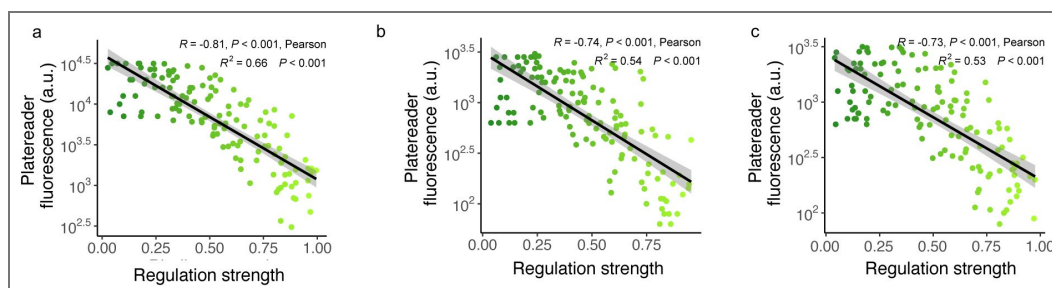
Supplementary Figure S10. Regulation strength distribution across the libraries.

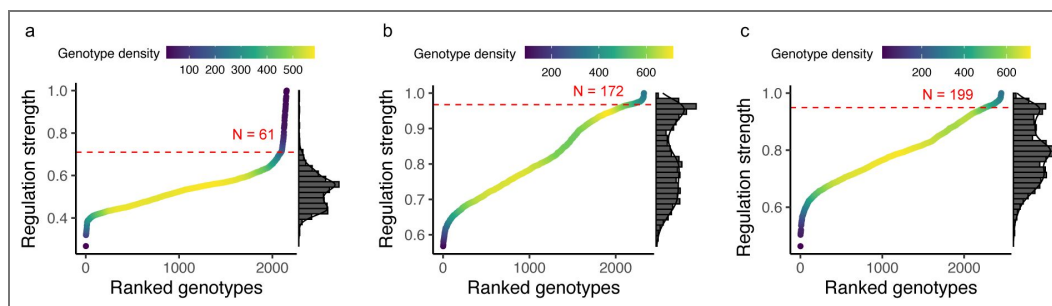
a. CRP library. The histogram illustrates the distribution of regulation strengths (normalized to the maximally observed value) among 31,975 library variants. The average regulation strength across all variants is highlighted by a dashed blue line (mean = 0.37), while the wild type's (WT = 0.71) strength is denoted by a dashed red line. A total of N=84 variants surpass the WT in regulation strength. **b. Fis library.** The histogram illustrates the distribution of regulation strengths (normalized to the maximum observed value) among 43,222 library variants. The average regulation strength across all variants is highlighted by a dashed blue line (mean = 0.57), while the wild type's (WT = 0.97) strength is denoted by a dashed red line. A total of N=172 variants surpass the WT in regulation strength. **c. IHF library.** The histogram illustrates the distribution of regulation strengths (normalized to the maximum observed value) among 41,325 library variants. The average regulation strength across all variants is highlighted by a dashed blue line (mean = 0.52), while the wild type's (WT = 0.95) strength is denoted by a dashed red line. A total of N=199 variants surpass the WT in regulation strength.



Supplementary Figure S11. Validating regulation strengths.

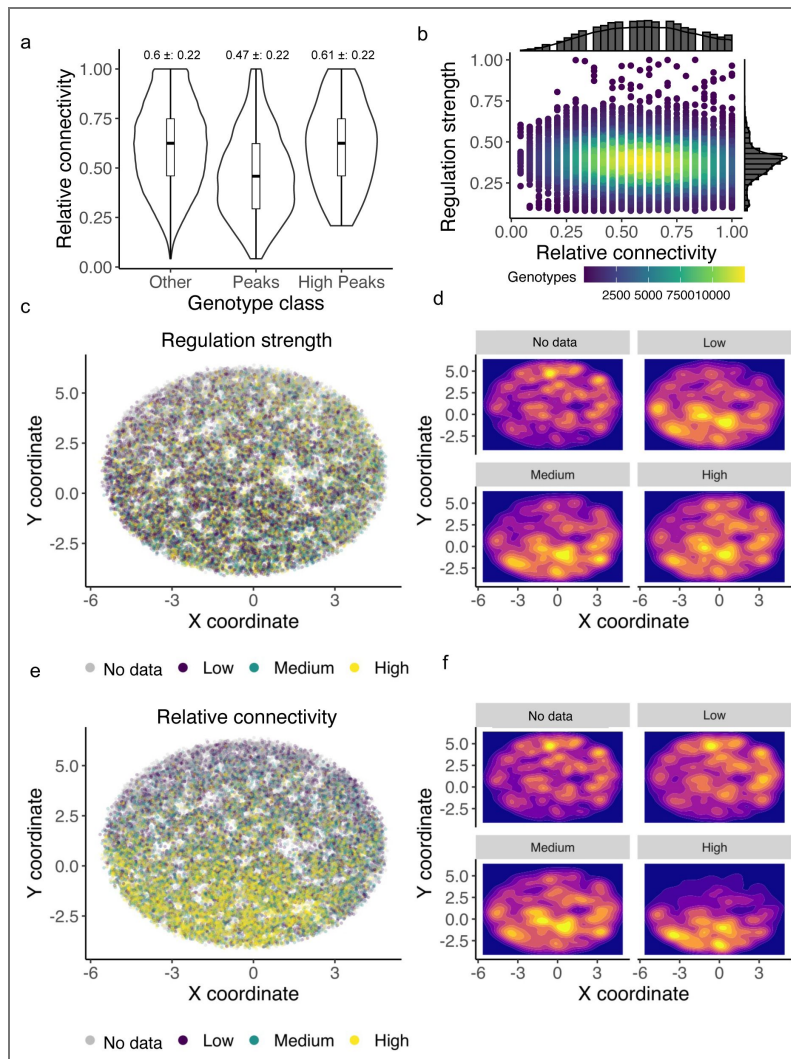
In this analysis, we validated the regulation strengths of 10 selected TFBSs for each bin and for each library (130 sequences per library). These TFBSs covered a wide range of regulation strengths determined from sort-seq data. The horizontal axis of each panel shows the regulation strengths of the 130 binding sites as quantified through sort-seq. The vertical axis shows the fluorescence levels of the same binding sites, as quantified by a plate reader (arbitrary units). The top of each panel shows the result of a test of the null hypothesis that the Pearson correlation coefficient R between the two quantities is equal to zero, together with a 95% confidence interval. The shaded grey area surrounding the regression line indicates the 95% confidence interval, i.e., the probability that the true linear regression line of the population lies within the confidence interval of the regression line calculated from the sample data at a confidence level of 95%. R^2 indicates the coefficient of determination. Each circle in the plot represents the mean of three independent (biological) replicate fluorescence measurements in a plate reader after eight hours of growth. Circle colors indicate GFP fluorescence from low (dark green) to high (light green). **a. Association between plate reader fluorescence and regulation strengths for the CRP library.** $R = -0.81$, $P < 0.001$; $R^2 = 0.66$, $P < 0.001$, $N = 130$. **b. Association between plate reader fluorescence and regulation strengths for the Fis library.** $R = -0.74$, $P < 0.001$; $R^2 = 0.54$, $P < 0.001$, $N = 130$. **c. Association between plate reader fluorescence and regulation strengths for the IHF library.** $R = -0.73$, $P < 0.001$; $R^2 = 0.53$, $P < 0.001$, $N = 130$.





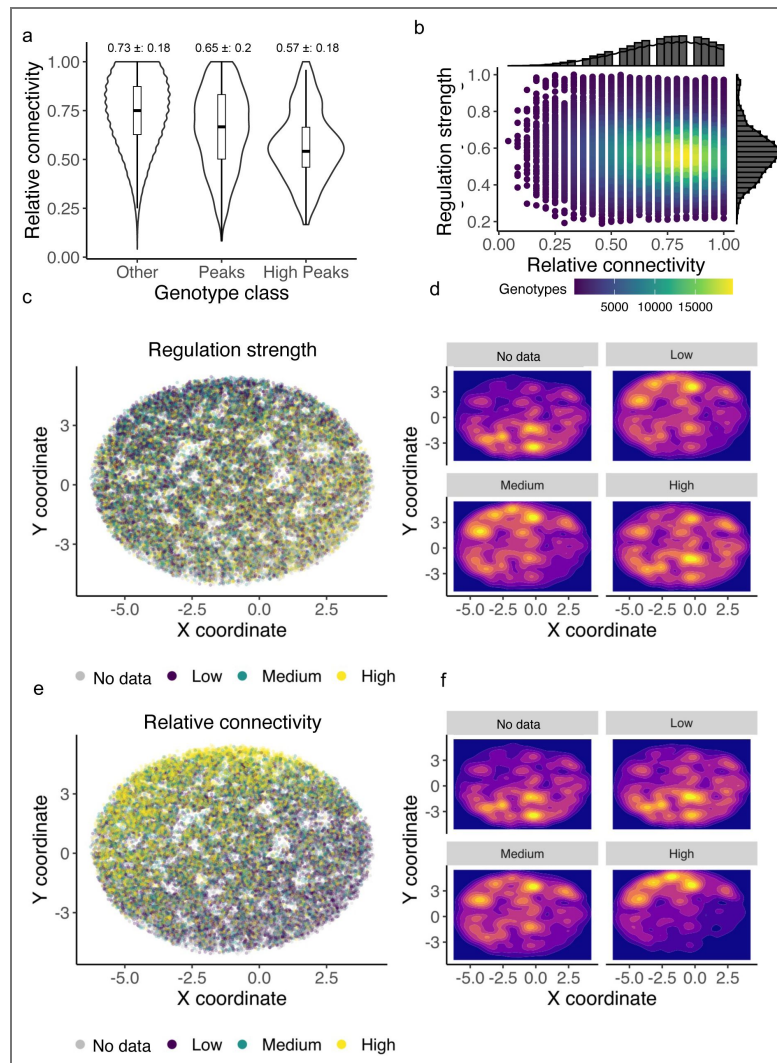
Supplementary Figure S12. Adaptive peaks and their regulation strengths across TF landscapes.

a. Adaptive peaks in the CRP landscape. The distribution of regulation strengths across the landscape's 2,154 adaptive peaks is shown. The WT's regulation strength (0.71) is marked by a horizontal dashed red line for reference. A total of N=61 peaks have higher regulation strength than the WT. The color gradient indicates genotype density along the regulation strength axis, with the histogram on the right detailing the regulation strength distribution. **b) like a) but for the Fis landscape,** based on its 2,312 adaptive peaks. (WT regulation strength: 0.97) N=172 peaks regulate expression more strongly than the WT. **c) like b) but for the IHF landscape,** based on its 2,434 adaptive peaks. (WT regulation strength: 0.95) N=199 peaks regulate expression more strongly than the WT.



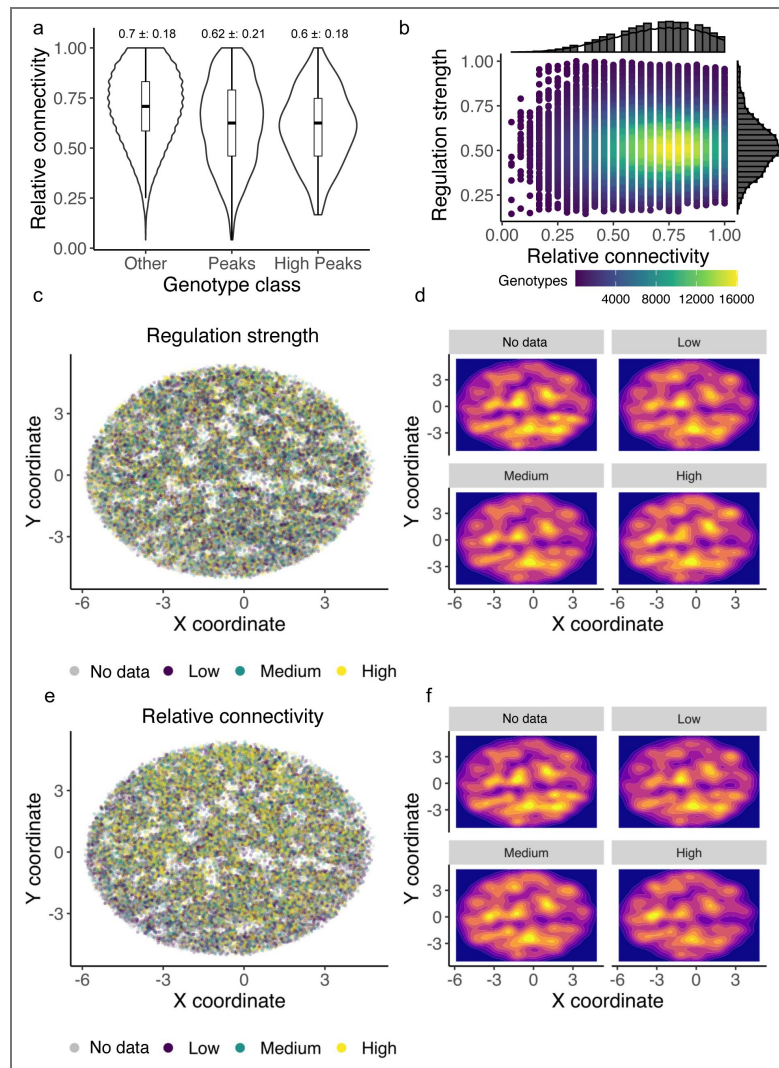
Supplementary Figure S13. Quantitative analysis of CRP landscape sparsity.

a. Distribution of relative connectivity among genotypes. Violin plots augmented with embedded boxplots illustrate the distribution of relative connectivity, i.e., the ratio of the empirically observed number of adjacent genotypes to the theoretical maximum number of $24(=8 \times 3)$ within a fully connected network. Genotypes are stratified into the three categories “other” (non-peak genotypes), “peaks” (excluding high peaks), and “high peaks”. Mean relative connectivities and standard deviations, are shown above each categorical division, revealing lower relative connectivity in peaks relative to non-peak genotypes (Two-sample Welch’s t-test, p -value = 4.59×10^{-77} , $N_1 = 29,821$; $N_2 = 2,093$). High peaks exhibit comparable or superior connectivity relative to non-peak genotypes (Two-sample Welch’s t-test, p -value = 1, $N_1 = 29,821$; $N_2 = 61$). **b. Weak association between connectivity and regulation strength.** Scatterplot of relative connectivity against regulation strength. The density of genotypes (circles) in the scatterplot area is represented by a color gradient from purple (low density) to yellow (high density). The weak association (Pearson correlation, $R = -0.13$; $t = -23.888$; degrees of freedom (df) = 31,973; p -value < 2.2×10^{-16}) indicates that strongly regulating genotypes are not necessarily densely connected. Marginal histograms adjacent to the scatterplot quantify the distributions of regulation strength and relative connectivity. **c. Visualization of CRP landscape based on regulation strength.** Visual overview of the CRP landscape with genotypes color-coded according to regulation strength (low strength: purple; medium: green; high: yellow; grey: genotypes with missing data). To enhance clarity, edges between genotypes are not shown. The landscape is projected onto a 2D space through a force-directed algorithm¹⁹⁷, with axes representing arbitrary units. **d. Spatial distribution of regulation strengths.** Each panel shows a density contour plot of the distribution of regulation strength scores of genotypes within one of four regulation strength categories (no data [missing data], low, medium, high). It indicates the density of genotypes within each of the four categories through a color gradient from blue (low density) to yellow (high density). **e. Visualization of CRP landscape based on relative connectivity.** Analogous to panel c, but for relative connectivity. Color-codes indicate relative connectivity (see color legend). **f. Spatial distribution of relative connectivity.** Like panel d, but for four categories of relative connectivity (no data, low, medium, high).



Supplementary Figure S14. Quantitative analysis of the Fis landscape sparsity.

a. Distribution of relative connectivity among genotypes. Violin plots augmented with embedded boxplots illustrate the distribution of relative connectivity, i.e., the ratio of the empirically observed number of adjacent genotypes to the theoretical maximum number of 24 ($=8 \times 3$) within a fully connected network. Genotypes are stratified into the three categories “other” (non-peak genotypes), “peaks” (excluding high peaks), and “high peaks.” Mean relative connectivities and standard deviations are shown above each categorical division, revealing lower relative connectivity in peaks relative to non-peak genotypes (Two-sample Welch’s t-test, $p\text{-value} = 4.59 \times 10^{-77}$, $N_1 = 40,910$; $N_2 = 2,312$). High peaks exhibit comparable or superior connectivity relative to non-peak genotypes (Two-sample Welch’s t-test, $p\text{-value} = 9.31 \times 10^{-13}$, $N_1 = 40,910$; $N_2 = 172$). **b. Weak association between connectivity and regulation strength.** Scatterplot of relative connectivity against regulation strength. The density of genotypes (circles) in the scatterplot area is represented by a color gradient from purple (low density) to yellow (high density). The weak association (Pearson correlation, $R = -0.13$; $t = -23.888$; degrees of freedom (df) = 31,973; $p\text{-value} < 2.2 \times 10^{-16}$) indicates that strongly regulating genotypes are not necessarily densely connected. Marginal histograms adjacent to the scatterplot quantify the distributions of regulation strength and relative connectivity. **c. Visualization of Fis landscape based on regulation strength.** Visual overview of the Fis landscape with genotypes color-coded according to regulation strength (low strength: purple; medium: green; high: yellow; grey: genotypes with missing data). To enhance clarity, edges between genotypes are not shown. The landscape is projected onto a 2D space through a force-directed algorithm¹⁹⁷, with axes representing arbitrary units. **d. Spatial distribution of regulation strengths.** Each panel shows a density contour plot of the distribution of regulation strength scores of genotypes within one of four regulation strength categories (no data [missing data], low, medium, high). The density of genotypes within each category is indicated by a color gradient from blue (low density) to yellow (high density). **e. Visualization of Fis landscape based on relative connectivity.** Analogous to panel c, but for relative connectivity. Color codes indicate relative connectivity (see color legend). **f. Spatial distribution of relative connectivity.** Similar to panel d, but for four categories of relative connectivity (no data, low, medium, high).

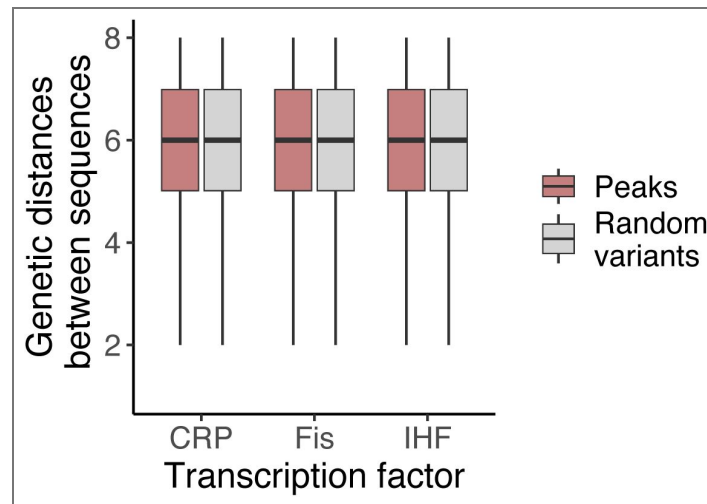


Supplementary Figure S15. Quantitative analysis of the IHF landscape sparsity.

a. Distribution of relative connectivity among genotypes. Violin plots augmented with embedded boxplots illustrate the distribution of relative connectivity, i.e., the ratio of the empirically observed number of adjacent genotypes to the theoretical maximum number of 24 ($=8 \times 3$) within a fully connected network. Genotypes are stratified into the three categories “other” (non-peak genotypes), “peaks” (excluding high peaks), and “high peaks.” Mean relative connectivities and standard deviations are shown above each categorical division, revealing lower relative connectivity in peaks relative to non-peak genotypes (Two-sample Welch’s t-test, p -value = 1.14×10^{-77} , $N_1 = 38,872$; $N_2 = 2,254$). High peaks exhibit comparable or superior connectivity relative to non-peak genotypes (Two-sample Welch’s t-test, p -value = 9.31×10^{-13} , $N_1 = 38,872$; $N_2 = 199$). **b. Weak association between connectivity and regulation strength.** Scatterplot of relative connectivity against regulation strength. The density of genotypes (circles) in the scatterplot area is represented by a color gradient from purple (low density) to yellow (high density). The slight positive correlation (Pearson correlation, $R = 0.017$; $t = 3.51$; degrees of freedom (df) = 41,323; p -value = 0.0004485) indicates that genotypes with elevated regulation strength are only marginally more likely to be densely connected. Marginal histograms adjacent to the scatterplot quantify the distributions of regulation strength and relative connectivity. **c. Visualization of IHF landscape based on regulation strength.** Visual overview of the IHF landscape with genotypes color-coded according to regulation strength (low strength: purple; medium: green; high: yellow; grey: genotypes with missing data). To enhance clarity, edges between genotypes are not shown. The landscape is projected onto a 2D space through a force-directed algorithm¹⁹⁷, with axes representing arbitrary units. **d. Spatial distribution of regulation strengths.** Each panel shows a density contour plot of the distribution of regulation strength scores of genotypes within one of four regulation strength categories (no data [missing data], low, medium, high). The density of genotypes within each category is indicated by a color gradient from blue (low density) to yellow (high density). **e. Visualization of IHF landscape based on relative connectivity.** Analogous to panel c, but for relative connectivity. Color codes indicate relative connectivity (see color legend). **f. Spatial distribution of relative connectivity.** Similar to panel d, but for four categories of relative connectivity (no data, low, medium, high).

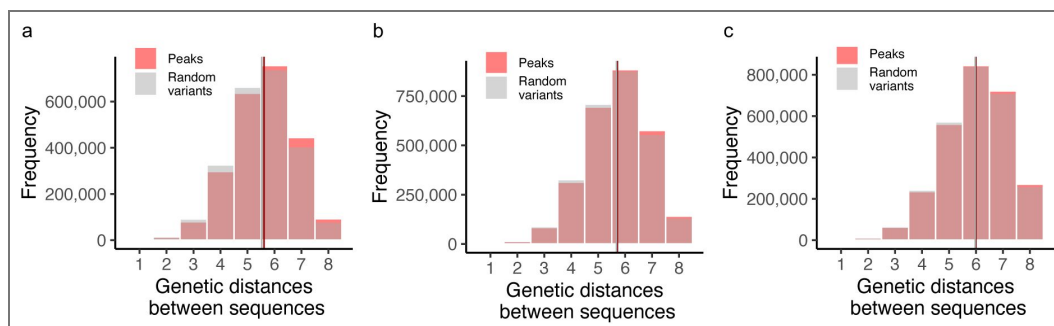
Supplementary Figure S16. Peaks are spread across each landscape.

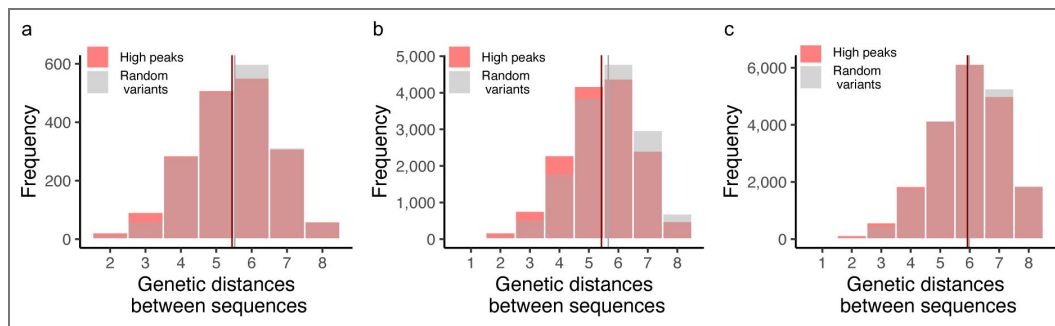
The vertical axis shows the genetic distance d between peaks and an equal number of randomly chosen non-peak genotypes (color legend) for the CRP ($N = 2,154$), Fis ($N = 2,312$), and IHF ($N = 2,434$) landscape (horizontal axis). (CRP: $d_{\text{peak}} = 6$, $d_{\text{random}} = 5.94$, $t\text{-statistic} = 52.49$, $p\text{-value} < 0.01$; Fis: $d_{\text{peak}} = 6$, $d_{\text{random}} = 5.97$, $t\text{-statistic} = 30.96$, $p\text{-value} < 0.01$; IHF: $d_{\text{peak}} = 6$, $d_{\text{random}} = 5.97$, $t\text{-statistic} = 34.07$, $p\text{-value} < 0.01$; null-hypothesis: means do not differ). Although the differences are statistically significant, peaks are almost as dispersed within genotype space as non-peaks. Each box covers the range between the first and third quartiles (IQR). The horizontal line within the box represents the median value, and whiskers span 1.5 times the IQR.



Supplementary Figure S17. Genetic diversity among peaks across TF landscapes.

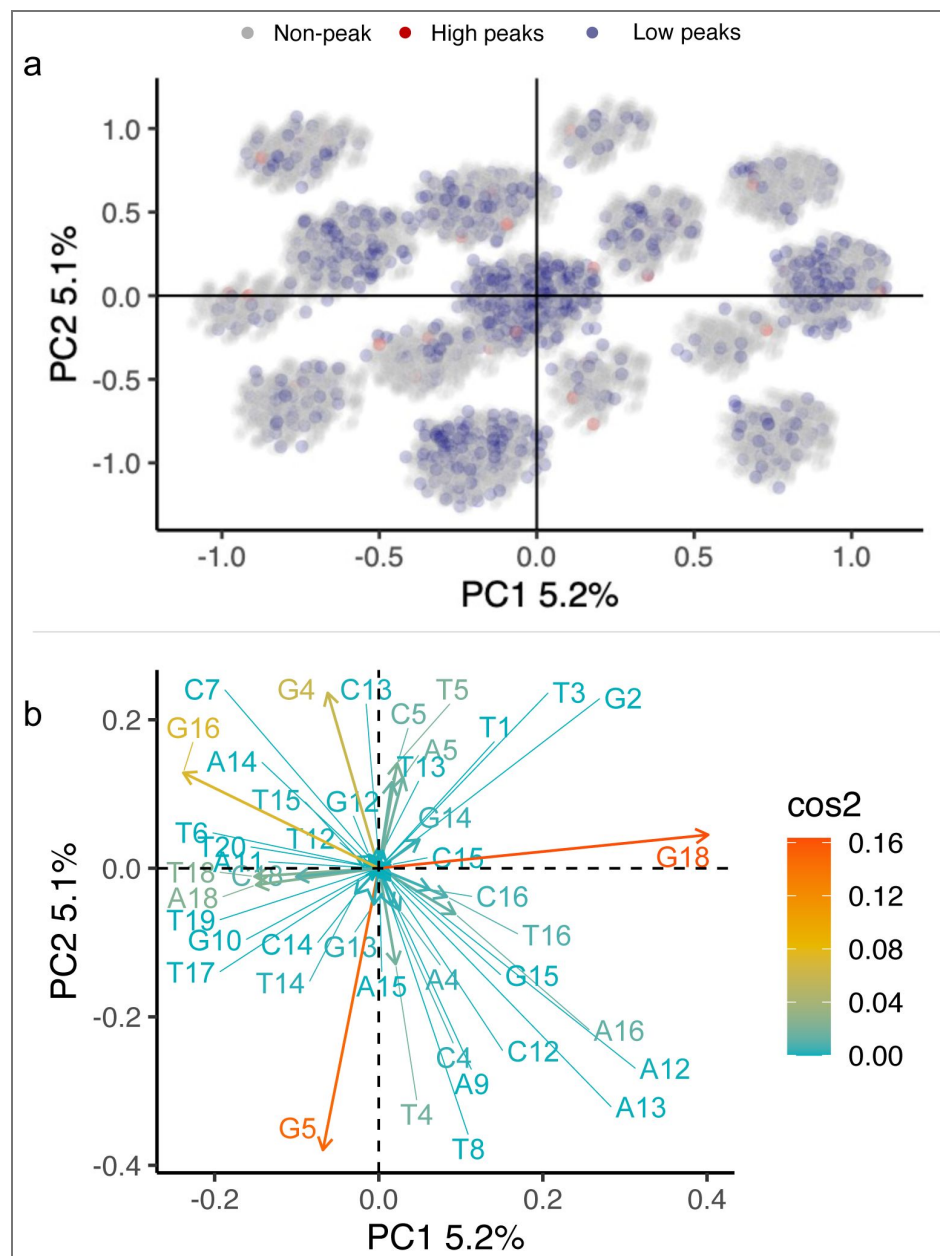
a. CRP landscape. Comparative analysis of genetic distances among nucleotide sequences from 2,154 peaks (red) versus 2,154 non-peak, random variants (grey). Mean genetic distances are displayed with vertical lines, showing a slight difference between peaks ($d = 5.6$, in red) and non-peaks ($d = 5.5$, in grey), demonstrating large genetic diversity within the peaks (Welch Two Sample t-test, $t = 78.299$, $df = 4637305$, $p\text{-value} < 2.2 \times 10^{-16}$). **b. Fis landscape.** Comparative analysis of genetic distances among nucleotide sequences from 2,312 peaks (red) versus 2,312 non-peak, random variants (grey). Mean genetic distances are displayed with vertical lines, showing no significant difference between peaks ($d = 5.7$, in red) and non-peaks ($d = 5.7$, in grey), demonstrating large genetic diversity within the peaks ($t = 27.234$, $df = 5412460$, $p\text{-value} < 2.2 \times 10^{-16}$). **c. IHF landscape.** Comparative analysis of genetic distances among nucleotide sequences from 2,434 peaks (red) versus 2,434 non-peak, random variants (grey). Mean genetic distances are displayed with vertical lines, showing no significant difference between peaks ($d = 6$, in red) and non-peaks ($d = 6$, in grey), demonstrating large genetic diversity within the peaks ($t = 54.689$, $df = 6014743$, $p\text{-value} < 2.2 \times 10^{-16}$).





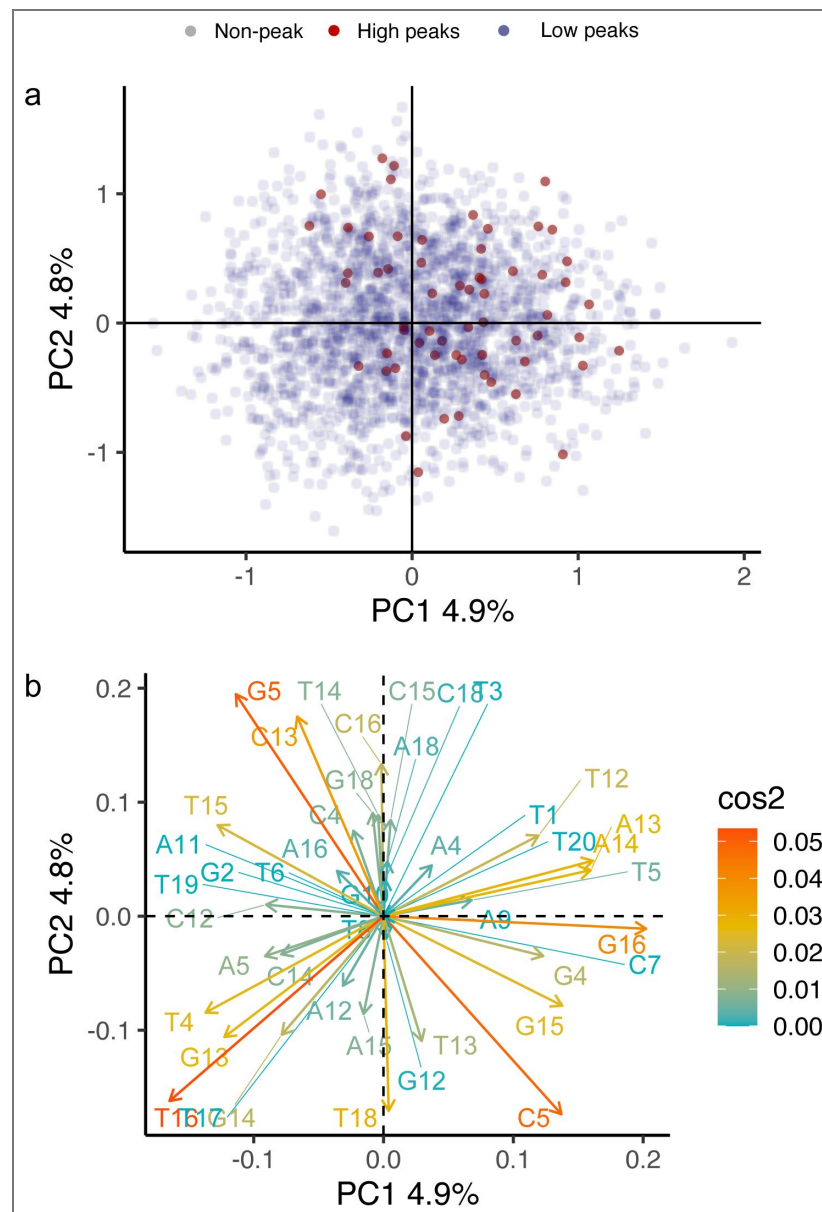
Supplementary Figure S18. Genetic diversity among high peaks across TF landscapes.

a. CRP landscape. The overlaid histograms compare the genetic diversity between 61 high regulation strength peaks (red) and 61 non-peak, randomly selected variants (grey). No significant difference exists (W) between high-regulation strength peaks (mean $d=5.4$, in red) compared to randomly selected genotypes (mean $d=5.5$, in grey; Welch Two Sample t-test, $t = -0.78555$, $df = 3653.7$, $p\text{-value} = 0.4322$). **c. Fis landscape.** The overlaid histograms compare the genetic diversity between 172 high-regulation strength peaks (red) and 172 non-peak, randomly selected genotypes (grey). No significant difference exists between high-regulation strength peaks (mean $d=5.4$, in red) and random genotypes (mean $d=5.8$, in grey; Welch Two Sample t-test, $t = -16.634$, $df = 29357$, $p\text{-value} < 2.2 \times 10^{-16}$). **c. IHF landscape.** The overlaid histograms compare the genetic diversity between 199 high regulation strength peaks (red) and 199 non-peak, randomly selected genotypes (grey). The analysis reveals no significant difference between high regulation strength peaks (mean $d=5.8$, in red) compared to the random variants (mean $d=5.8$, in grey; Welch Two Sample t-test, $t = 0.73365$, $df = 39366$, $p\text{-value} = 0.4632$).



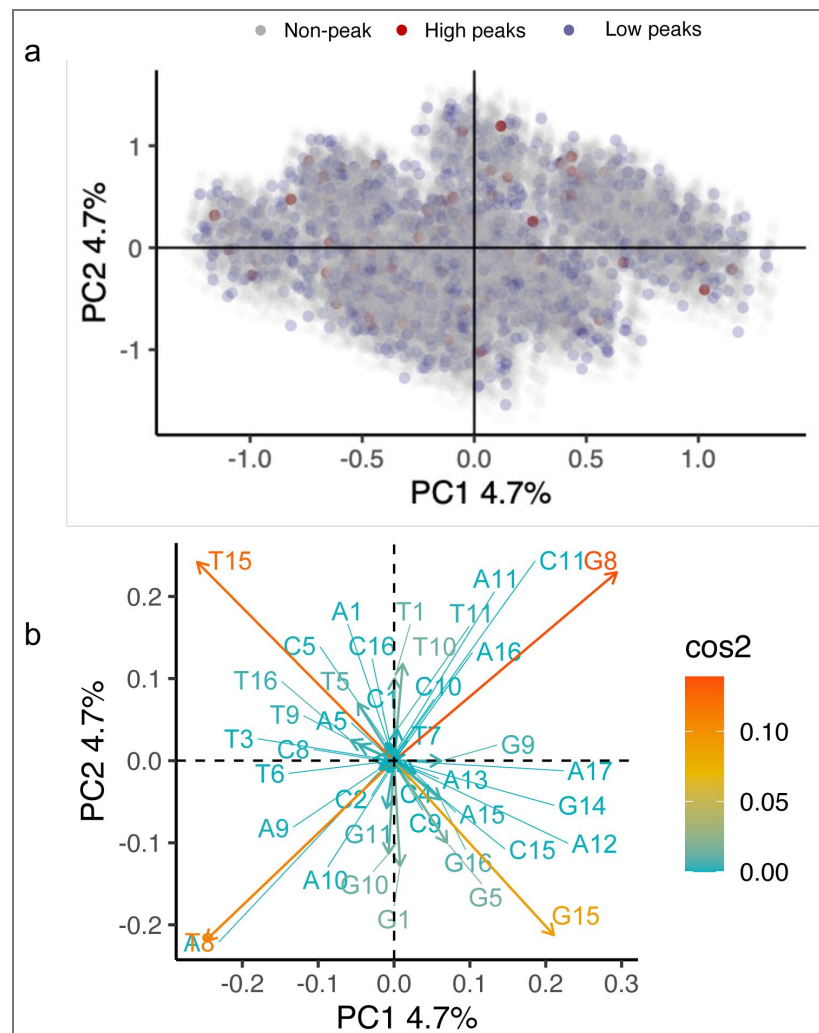
Supplementary Figure S19. Principal component analysis of the CRP landscape.

a. PCA plot. A principal component analysis reveals that adaptive peaks are distributed widely within the genotype space (Supplementary Methods). The panels show principal components (PC) 1 and 2. Axes labels also indicate the percentage of variation explained by each component. Every circle corresponds to one among the 31,975 genotypes within the landscape. Light blue and light red circles represent low and high peaks, respectively, while grey circles represent non-peak variants. The clustered structure in the genotype space is due to the different influence of individual nucleotides on the variation observed in the data (see panel b). **b. The contribution of individual nucleotides to principal component analysis.** A PCA contribution plot is a way to visualize the relative importance of different variables to the variation observed in the data. In this plot, the contribution of each variable (nucleotide identity at each position of a TFBS) is expressed as the square of the cosine of the angle ($\cos^2$) between the variable's vector (column representing the presence/absence of a base letter at each position in the sequence) and each principal component axis. This quantity is represented as an arrow that indicates the correlation of the variable with PC1 and PC2, the two principal components that capture the largest amount of variation in the data. Both length and color of the arrow represent the contribution of the variable to the variation observed in the data. A high $\cos^2$ value (red) indicates that the variable is strongly correlated with the principal component, and therefore makes a large contribution to the variation observed in the data. A low $\cos^2$ value (blue) indicates that the variable is weakly correlated with the principal component, and therefore makes a small contribution to the variation observed in the data. Each arrow's length represents the importance of the variable's contribution relative to other variables in the plot. Each nucleotide is represented with a base letter (A, T, C, G) followed by a number that indicates the position of that base in the binding site sequence (e.g. G18 stands for a guanine at position 18 of the binding site).



Supplementary Figure S20. Principal component analysis of the Fis landscape.

a. PCA plot. A principal component analysis reveals that adaptive peaks are distributed widely within the genotype space (Supplementary Methods). The panels show principal components (PC) 1 and 2. Axes labels also indicate the percentage of variation explained by each component. Every circle corresponds to one among the 43,222 genotypes within the landscape. Light blue and light red circles represent low and high peaks, respectively, while grey circles represent non-peak variants. The clustered structure of the data in genotype space is due to the different influence of individual nucleotides on the variation observed in the data (see panel b). **b. The contribution of individual nucleotides to principal component analysis.** A PCA contribution plot visualizes the relative importance of different variables to the variation observed in the data. In this plot, the contribution of each variable (nucleotide identity at each position of a TFBS) is expressed as the square of the cosine of the angle ($\cos^2$) between the variable's vector (column representing the presence/absence of a base letter at each position in the sequence) and each principal component axis. This quantity is represented as an arrow that indicates the correlation of the variable with PC1 and PC2, the two principal components that capture the largest amount of variation in the data. Both the length and color of the arrow represent the contribution of the variable to the variation observed in the data. A high $\cos^2$ value (red) indicates that the variable is strongly correlated with the principal component and therefore makes a large contribution to the variation observed in the data. A low $\cos^2$ value (blue) indicates that the variable is weakly correlated with the principal component and therefore makes a small contribution to the variation observed in the data. Each arrow's length represents the importance of the variable's contribution relative to other variables in the plot. Each nucleotide is represented with a base letter (A, T, C, G) followed by a number that indicates the position of that base in the binding site sequence (e.g., G18 stands for a guanine at position 18 of the binding site).

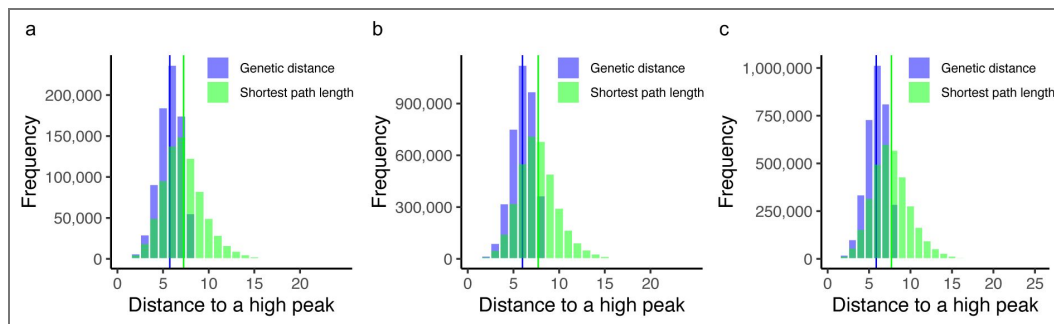


Supplementary Figure S21. Principal Component Analysis of the IHF Landscape.

a. PCA Plot. A principal component analysis reveals that adaptive peaks are distributed widely within the genotype space (Supplementary Methods). The panels show principal components (PC) 1 and 2. Axes labels also indicate the percentage of variation explained by each component. Every circle corresponds to one among the 41,325 genotypes within the landscape. Light blue and light red circles represent low and high peaks, respectively, while grey circles represent non-peak variants. The clustered structure in the genotype space is due to the different influence of individual nucleotides on the variation observed in the data (see panel b). **b. The contribution of individual nucleotides to principal component analysis.** A PCA contribution plot is a way to visualize the relative importance of different variables to the variation observed in the data. In this plot, the contribution of each variable (nucleotide identity at each position of a TFBS) is expressed as the square of the cosine of the angle ($\cos^2$) between the variable's vector (column representing the presence/absence of a base letter at each position in the sequence) and each principal component axis. This quantity is represented as an arrow that indicates the correlation of the variable with PC1 and PC2, the two principal components that capture the largest amount of variation in the data. Both the length and color of the arrow represent the contribution of the variable to the variation observed in the data. A high $\cos^2$ value (red) indicates that the variable is strongly correlated with the principal component and therefore makes a large contribution to the variation observed in the data. A low $\cos^2$ value (blue) indicates that the variable is weakly correlated with the principal component and therefore makes a small contribution to the variation observed in the data. Each arrow's length represents the importance of the variable's contribution relative to other variables in the plot. Each nucleotide is represented with a base letter (A, T, C, G) followed by a number that indicates the position of that base in the binding site sequence (e.g., G18 stands for a guanine at position 18 of the binding site).

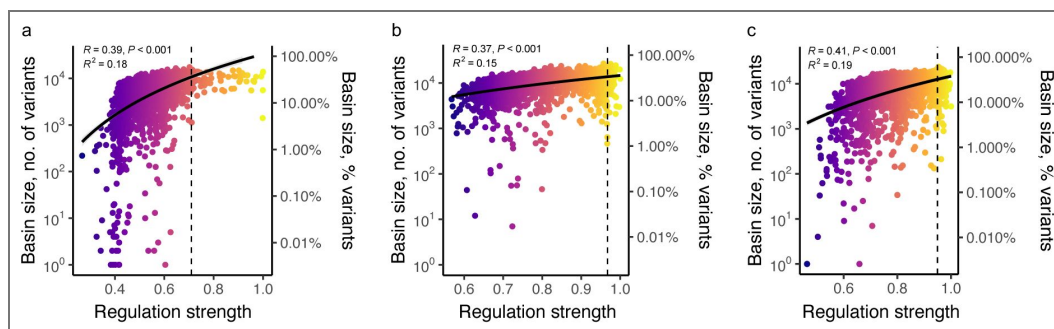
Supplementary Figure S22. Accessible paths to high peaks are less than two mutational steps longer than the shortest genetic distances to the peak.

Each composite histogram shows the distribution of genetic distances (blue) and shortest accessible path length (green) for all pairs of high peaks and non-peak genotypes for one of the three landscapes. We used a t-test of the null hypothesis that the means of these distance distributions are statistically indistinguishable. This null hypothesis is rejected for all three landscapes. **a. CRP Landscape.** Mean genetic distance: 5.74; mean number of mutational steps in shortest accessible paths: 7.2; $t = 500.17$, $df = 1,212,712$, $p < 2.2 \times 10^{-16}$. **b. Fis Landscape.** Mean genetic distance: 5.98; mean number of mutational steps in shortest accessible paths: 7.7; $t = -1271$, $df = 5,639,001$, $p < 2.2 \times 10^{-16}$. **c. IHF Landscape.** Mean genetic distance: 5.78; mean number of mutational steps in shortest accessible paths: 7.6; $t = -1365$, $df = 6,636,001$, $p < 2.2 \times 10^{-16}$.



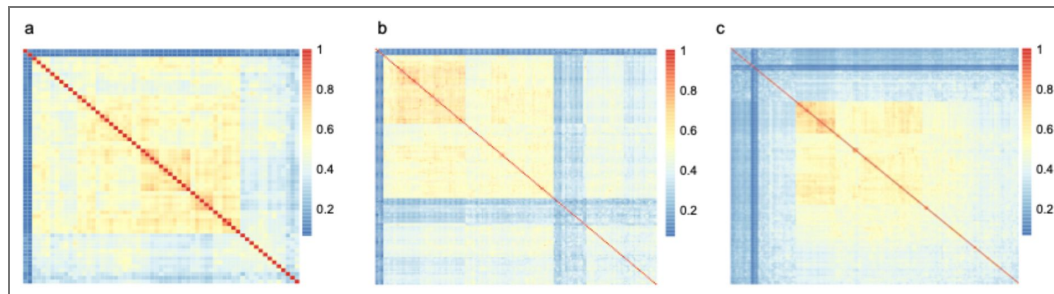
Supplementary Figure S23. Moderate positive correlation between peak regulation strength and basin size for the three regulatory landscapes.

a. CRP landscape. The scatter plot shows the association between regulation strength conveyed by a peak variant (horizontal axis, colors from dark blue [low] to yellow [high]) and the size of its basin of attraction (absolute numbers on the left vertical axis, percentages on the right vertical axis, note the logarithmic scales). The dashed vertical line represents the regulation strength of the wild type. The black curve represents a linear regression line (in linear space). R is the Pearson correlation coefficient ($R=0.39$) and R^2 is the coefficient of determination for the linear regression model ($R^2=0.18$, $N = 2,154$). **b. Fis landscape.** Like panel *a* ($R=0.37$, $R^2=0.15$, $N = 2,312$). **c. IHF landscape.** Like panel *a* ($R=0.41$, $R^2=0.19$, $N = 2,434$).



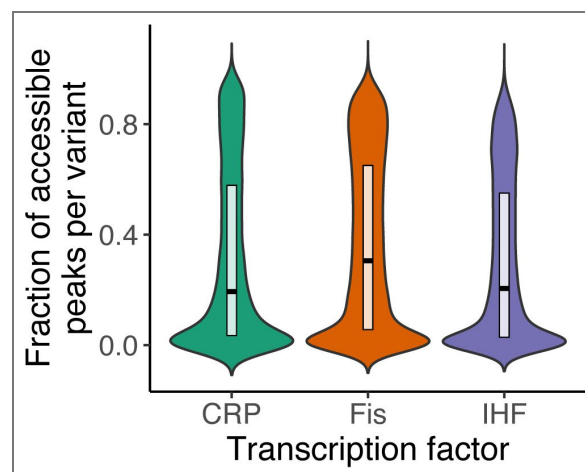
Supplementary Figure S24. The basins of attractions of high peaks overlap.

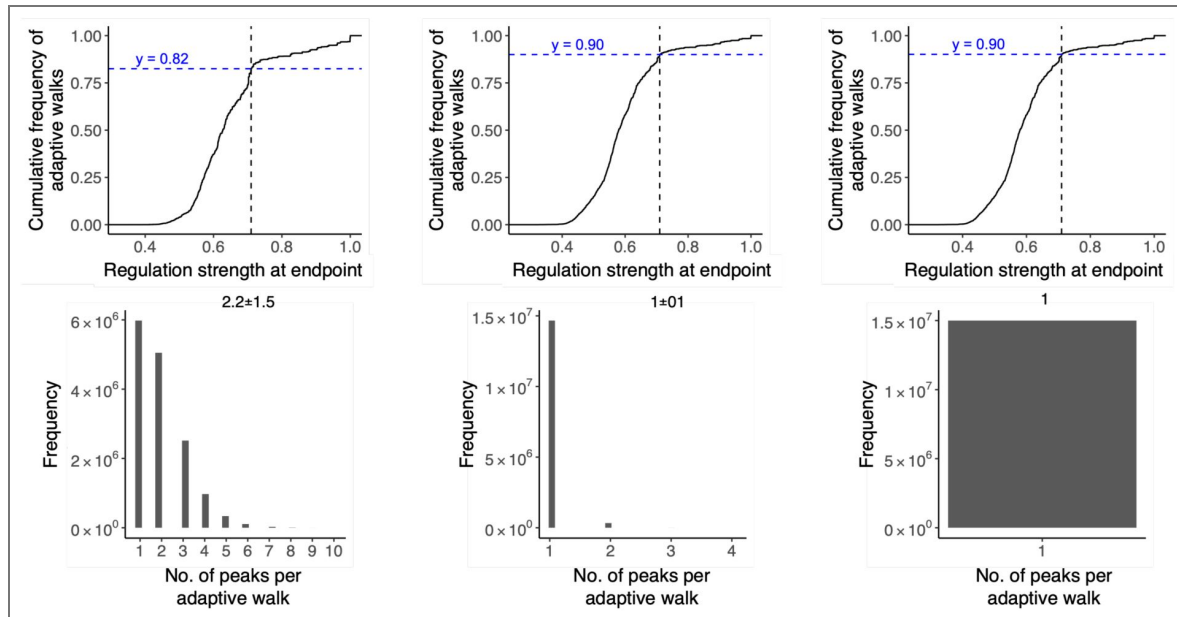
Each heatmap matrix quantifies the proportion of shared TFBS variants in the basins of attractions of all pairs of high peaks (Methods). Each matrix entry quantifies the basin overlap for one pair of high peaks using the Jaccard index (**Supplementary Methods 7.5**), where row and column indices correspond to specific pairs of peaks. The maximum overlap of one (red) signifies that two basins share all variants, while the minimum value of zero (blue) indicates that they share no variants. **a. Basin overlap in the CRP landscape.** The average Jaccard index is 0.461 with a standard deviation of ± 0.144 , based on 3,721 pairwise comparisons. **b. Basin overlap in the Fis landscape.** The average Jaccard index is 0.426 with a standard deviation of ± 0.136 , derived from 29,584 pairwise comparisons. **c. Basin overlap in the IHF landscape.** The average Jaccard index is 0.364, with a standard deviation of ± 0.145 , based on 39,601 pairwise comparisons.



Supplementary Figure S25. Distribution of the fraction of accessible high peaks per TFBS variant across TF landscapes.

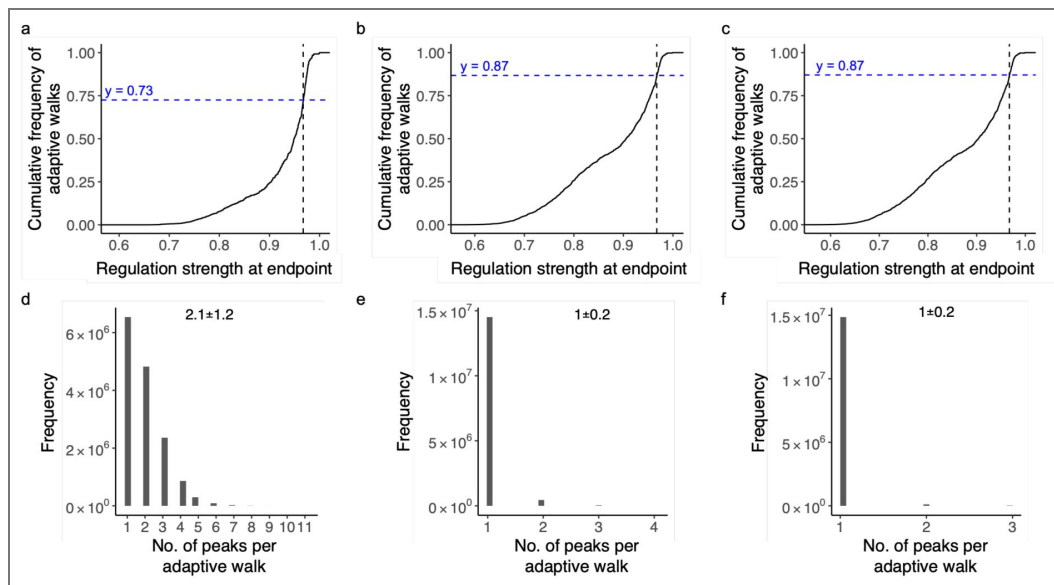
The figure shows violin plots with embedded boxplots, for the distribution of high peak accessibility of each variant within the CRP (green), Fis (orange), and IHF (purple) landscape. The y-axis quantifies peak accessibility as the fraction of the total number of high peaks that can be accessed from each (non-peak) genotype within each TF landscape, revealing the extent to which individual variants can potentially access different high peaks. In the CRP landscape, each genotype can access, on average, a proportion of 0.314 ± 0.315 (mean $\pm$ s.d.) high peaks from the total number of high peaks, with a median value of 0.194. Notably, 5,152 out of 29,821 genotypes (17%) can access no high peak. Fis: mean high peak accessibility: 0.363 ± 0.311 (median: 0.305), with 4,304 out of 40,895 (10%) genotypes lacking access to any high peaks. IHF: mean peak accessibility: 0.301 ± 0.289 (median: 0.205); 4,629 out of 38,872 genotypes (12%) show lack access to high peaks.





Supplementary Figure S26. Kimura adaptive walks on the CRP landscape.

a-c. Fraction of adaptive walks attaining TFBS variants with strong regulation for different population sizes. The vertical axis indicates the cumulative fraction of adaptive walks on the CRP landscape that end at a genotype with a regulation strength shown on the horizontal axes. Data is based on 15 million simulated adaptive walks (15,000 random starting genotypes $\times$ 1,000 adaptive walks each). The vertical dashed line indicates the regulation strength of the wild-type TFBS for CRP, and can be used to infer the proportion of walks that terminate at lower regulatory peaks (Horizontal line, blue lettering). For example, a value of $y=0.82$ implies that 82% of walks terminate at peaks lower than the wild-type. The remaining 18% reach higher peaks. **a. Population size 10^2 .** 18% of peaks terminate at peaks higher than the wild-type. **b. Population size 10^5 .** 10% of the walks terminate at higher peaks. **c. Population size 10^8 .** 10% of walks terminate at higher peaks. **d-f. Number of high peaks reached by a single adaptive walks.** The histograms show the distribution of the number of high peaks visited by a single adaptive walk, based on the same 15×10^6 adaptive walks analyzed in a-c. Note that a walk can visit multiple peaks when genetic drift is strong, because genetic drift can displace a population from an adaptive peak. The mean $\pm$ s.d. of the distribution is shown at the top of each plot. Population sizes: **d** 10^2 , **e** 10^5 , and **f** 10^8 .

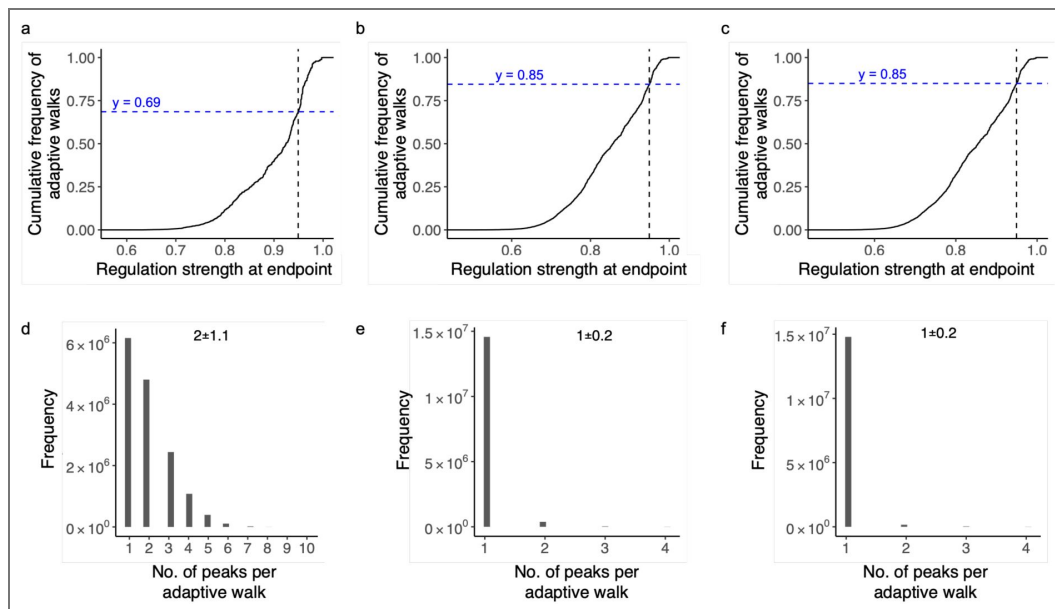


Supplementary Figure S27. Kimura adaptive walks on the Fis landscape.

a-c. Fraction of adaptive walks attaining high peaks for different population sizes. The vertical axis indicates the cumulative percentage of adaptive walks on the Fis landscape that end at a genotype with a regulation strength shown on the horizontal axis. Data is based on 15 million simulated adaptive walks (15,000 starting genotypes $\times$ 1,000 adaptive walks each). The vertical dashed line indicates the regulation strength of the wild-type TFBS for Fis, and can be used to infer the proportion of walks that terminate at lower regulatory peaks (horizontal line, blue lettering). For example, a y-value of 0.9 implies that 90% of walks terminate at peaks lower than the wild-type, with the remaining 10% reaching higher peaks. **a. Population size 10^2 .** 27% of walks terminate at peaks higher than the wild-type. **b. Population size 10^5 .** 13% of walks terminate at higher peaks. **c. Population size 10^8 .** 13% of walks terminate at higher peaks. **d-f. Number of high peaks reached by individual adaptive walks.** The histograms show the distribution of the number of high peaks visited by a single adaptive walk, based on the same 150 million adaptive walks analyzed in a-c. Note that a walk can visit multiple peaks when genetic drift is strong because genetic drift can displace a population from an adaptive peak. The mean $\pm$ s.d. of the distribution is shown at the top of each plot. Population sizes: **d** 10^2 , **e** 10^5 , and **f** 10^8 .

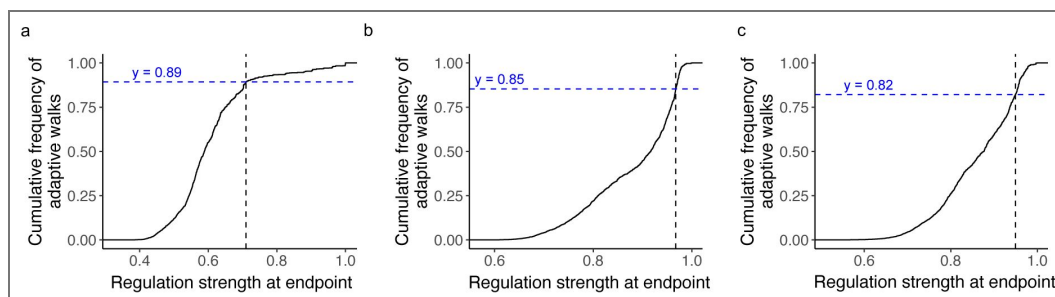
Supplementary Figure S28. Kimura adaptive walks on the IHF landscape.

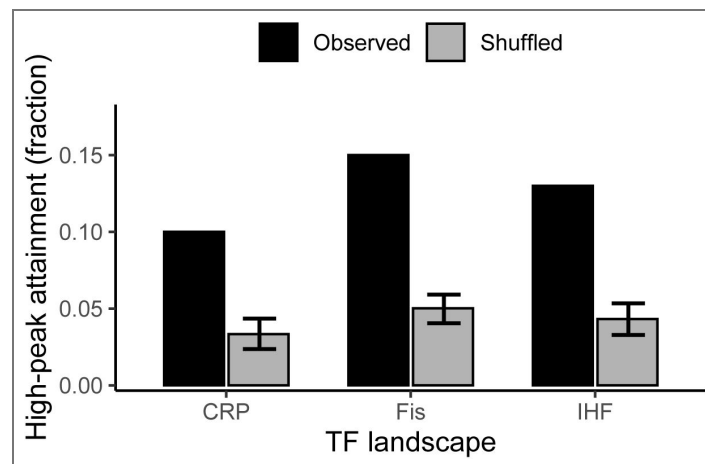
a-c. Fraction of adaptive walks attaining high peaks for different population sizes. The vertical axis indicates the cumulative percentage of adaptive walks on the IHF landscape that end at a genotype with a regulation strength shown on the horizontal axis. Data is based on 15 million simulated adaptive walks (15,000 starting points $\times$ 1,000 adaptive walks each). The vertical dashed line indicates the regulation strength of the wild-type TFBS for IHF, and can be used to infer the proportion of walks that terminate at lower regulatory peaks (horizontal line, blue lettering). For example, a y-value of 0.9 implies that 90% of walks terminate at peaks lower than the wild-type, with the remaining 10% reaching higher peaks. **a. Population size 10^2 .** 31% of walks terminate at peaks higher than the wild-type. **b. Population size 10^5 .** 15% of walks terminate at higher peaks. **c. Population size 10^8 .** 15% of walks terminate at higher peaks. **d-f. Number of high peaks reached by individual adaptive walks.** The histograms show the distribution of the number of high peaks visited by a single adaptive walk, based on the same 150 million adaptive walks analyzed in a-c. Note that a walk can visit multiple peaks when genetic drift is strong because genetic drift can displace a population from an adaptive peak. The mean $\pm$ s.d. of the distribution is shown at the top of each plot. Population sizes: **d** 10^2 , **e** 10^5 , and **f** 10^8 .



Supplementary Figure S29. Fraction of greedy adaptive walks attaining high peaks for different TFs.

a-c. Fraction of greedy adaptive walks attaining high peaks for different TFs. The vertical axis indicates the cumulative distribution function (CDF) of adaptive walks on each TF landscape that end at a genotype with a regulation strength shown on the horizontal axis. Data is based on 15,000 adaptive walks (15,000 starting points $\times$ 1 adaptive greedy walk for each) for each TF landscape. The vertical dashed line indicates the regulation strength of the wild-type TFBS for their respective TFs and can be used to infer the proportion of walks that terminate at lower regulatory peaks (horizontal line, blue lettering). For example, a value of $y=0.9$ implies that 90% of walks terminate at peaks lower than the wild-type, with the remaining 10% terminating at higher peaks. **a. CDF for adaptive walks in the CRP landscape.** Note the intersection at $y=0.89$, meaning that 11% of the walks terminated on high peaks. **b. CDF for adaptive walks in the Fis landscape.** Note the intersection at $y=0.85$, meaning that 15% of the walks terminated on high peaks. **c. CDF for adaptive walks in the IHF landscape.** Note the intersection at $y=0.82$, meaning that 18% of the walks terminated on high peaks.



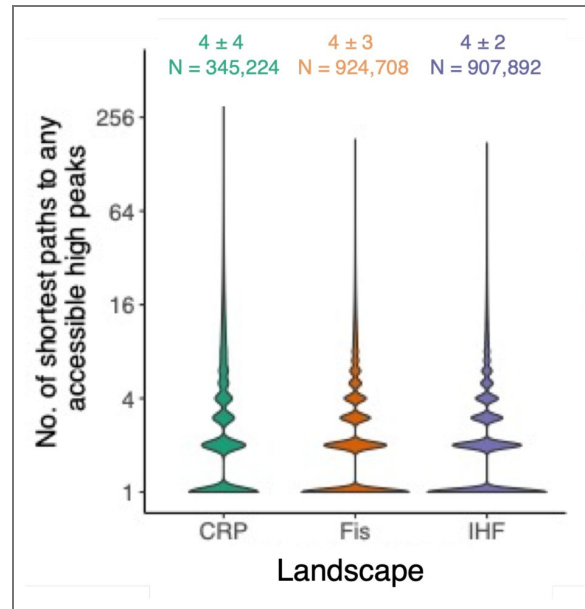


Supplementary Figure S30. Empirical regulatory landscapes enable more frequent access to strong regulatory peaks than randomized landscapes.

Bars show the fraction of adaptive walks that reach a high regulatory peak in empirical landscapes (black) and in randomized null landscapes (gray) for the transcription factors CRP, Fis, and IHF. We define high peaks as local peaks whose regulation strength exceeds that of the wild-type TFBS for the corresponding TF. We modeled adaptive evolution using Kimura walks for a population size of 10^8 . For each landscape, we simulated 10^3 adaptive walks starting from each of 15,000 randomly and uniformly selected non-peak TFBS genotypes, allowing up to 8 mutational fixation steps per walk until a peak is reached. We quantified the null expectations from 10^3 shuffled landscapes per TF by permuting regulation-strength values across genotypes while preserving the sampled genotype network (see **Supplementary Methods 9**). Gray bars show the mean fraction of walks reaching high peaks across shuffled landscapes; error bars indicate $\pm$ s.d. For all three TFs, empirical landscapes show significantly higher accessibility of high regulatory peaks than expected under the null model, with empirical landscapes exhibiting on average a 3-fold higher accessibility of high regulatory peaks (One-sided Monte Carlo permutation test; $N_2 = 1$ empirical landscape, $N_9 = 1,000$ shuffled landscapes; CRP: observed = 0.10, null = 0.035 ± 0.010 , $z = 13.4$, $p < 0.001$; Fis: observed = 0.15, null = 0.050 ± 0.010 , $z = 20.9$, $p < 0.001$; IHF: observed = 0.13, null = 0.045 ± 0.010 , $z = 16.8$, $p < 0.001$; $N_{\text{walks}} = 1.5 \times 10^3$ per landscape).

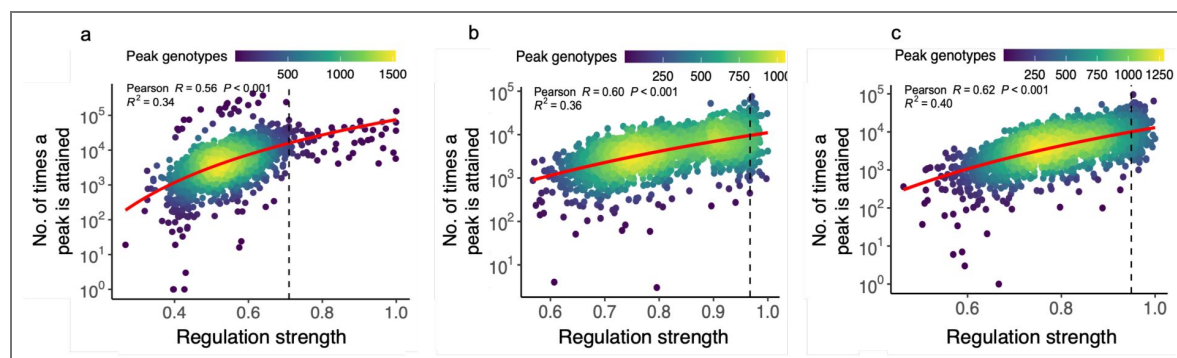
Supplementary Figure S31. Diversity of alternative shortest paths for pairs of starting genotypes and their respective attained peaks.

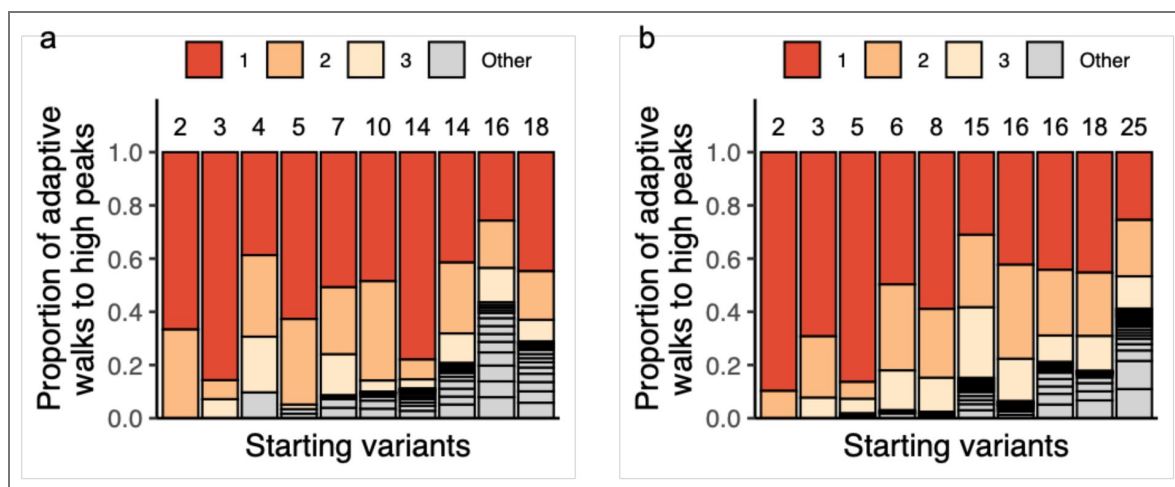
Each violin plot is based on data from 15,000 randomly (and uniformly) sampled starting (non-peak) genotypes from each TF landscape – CRP, Fis and IHF (see color legend) – and on 1'000 Kimura adaptive walks per starting genotypes. For each pair of a starting genotype and a high peak that is attainable from this starting genotype, we calculated the number of distinct shortest paths reaching the peak (vertical axis, note logarithmic scale). The width of the violin plot at any given y-axis value represents the density of data points at that value. Wider sections indicate a higher density of data points. The text at the top of each violin indicates the median number of shortest paths and the interquartile range (IQR) for each TF. For example, in the CRP landscape peaks are reached via a median number ($\pm$ IQR) of 4 ($\pm$ 4) shortest paths, with a sample size (number of starting genotypes $\times$ attainable peak pairs) of $N = 345,224$. The violin regions appear discretized because the number of shortest paths is a discrete variable, not a continuous one. This means that the data can only take on specific integer values, resulting in discrete regions within the violin plots with a non-zero density of data points.



Supplementary Figure S32. Peak genotypes conveying stronger regulation are also reached more often during adaptive evolution.

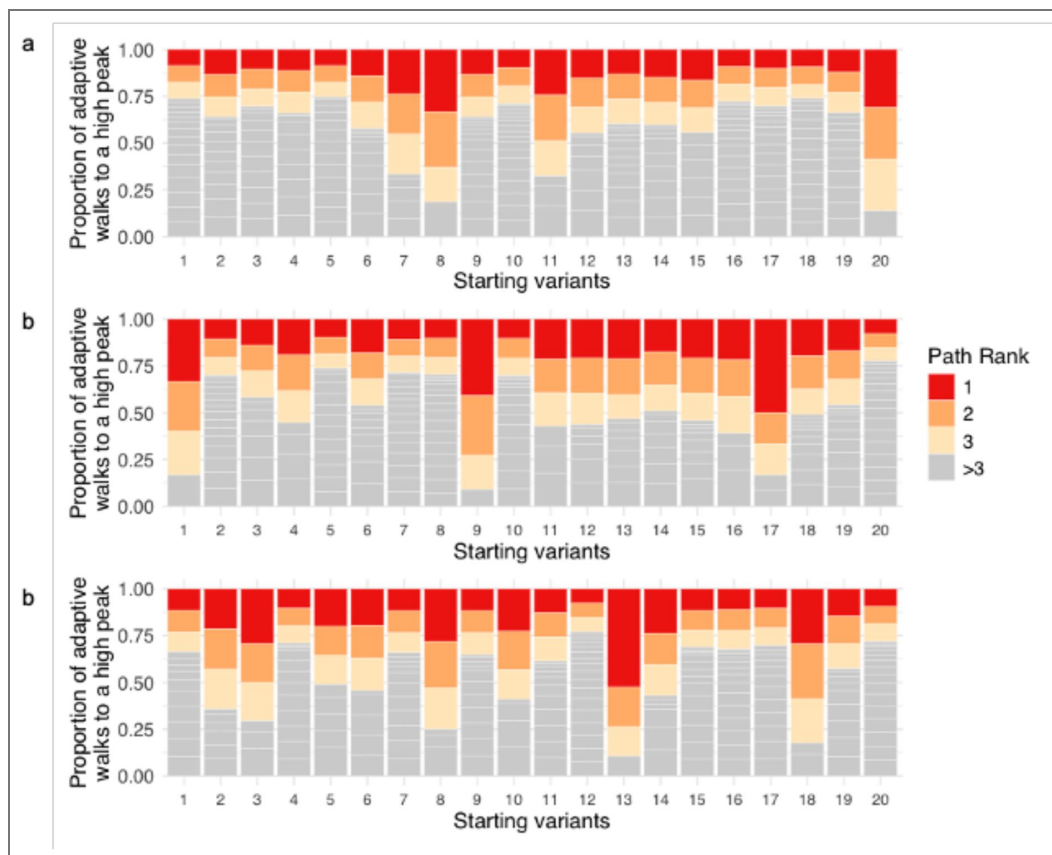
Each scatter plot represents the association between the strength of regulation conveyed by a peak genotype (horizontal axis) and the number of times the peak is attained (vertical axis, note the logarithmic scale). The data for each plot is based on 15,000 randomly (and uniformly) sampled starting (non-peak) genotypes from each TF landscape – CRP, Fis and IHF (see color legend) – and on 1'000 Kimura adaptive walks per starting genotypes (15×10^6 adaptive walks per landscape). The heatmap colors represent the density of peak genotypes from low (dark blue) to high (yellow). The dashed vertical line represents the regulation strength of the wild type TFBS. The red curve is derived from linear regression (in linear space). P-values in each panel result from a test of the null hypothesis that the two variables are not associated (Pearson correlation coefficient $R=0$; Coefficient of determination $R^2=0$). **a. CRP landscape** ($R=0.56$; $R^2=0.34$, $N = 2,154$). **b. Fis landscape** ($R=0.6$, $R^2=0.36$, $N = 2,312$). **c. IHF landscape** ($R=0.62$, $R^2=0.4$, $N = 2,434$).





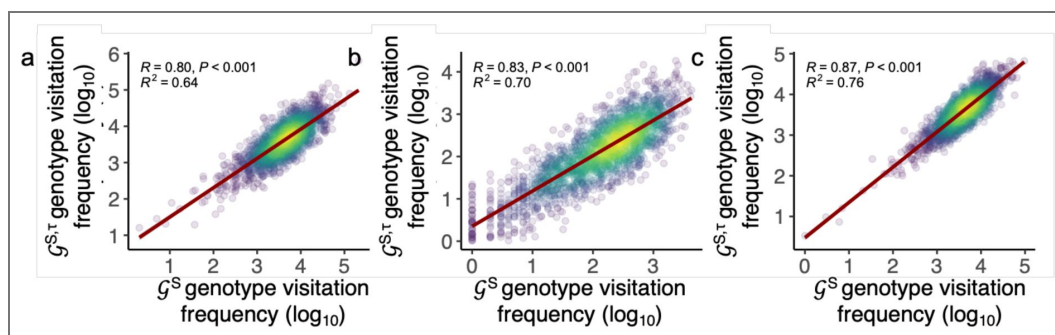
Supplementary Figures S33. The frequency distribution of attained high peaks is biased towards some frequently attained peaks.

We randomly and uniformly sampled 10 starting genotypes, started 10^3 adaptive walks from each, and recorded the number and frequency of distinct high peaks attained in these random walks. Each starting variant is symbolized by a vertical bar. The number of stacks within each bar (delineated by horizontal lines, also indicated by an integer above each bar) shows the number of high peaks reached by the 10^3 adaptive walks. Starting variants are ordered in ascending order based on this number of attained peaks. Stack height indicates the fraction of walks that reached the same peak, and is indicated in red, orange, and yellow for the three most frequently attained peaks. **a. Fis landscape. b. IHF landscape.**



Supplementary Figure S34. The frequency distribution of traversed paths to high peaks is biased towards some paths that are frequently traversed.

We randomly and uniformly sampled 20 starting genotypes, initiated 10^3 adaptive walks from each, and recorded the number and frequency of distinct paths traversed by these random walks for each attained peak. We then randomly and uniformly sampled a high peak attained by each starting variant for our analysis. Each starting variant is represented by a vertical bar. The number of stacks within each bar indicates the number of unique paths traversed to reach the sampled high peak. Stack height represents the fraction of walks that reached the high peak, with the three most frequently traversed paths in red, orange, and yellow. **a. CRP landscape. b. Fis landscape. c. IHF landscape.**



Supplementary Figure S35. Evolutionary accessibility is robust to experimental uncertainty in regulation strength.

Each panel shows a pairwise comparison of the log-transformed genotype visitation frequency during adaptive (Kimura) walks in the absence of noise (G^S , x-axis) and after allowing for empirically estimated measurement noise in regulation strengths ($G^{S,T}$, y-axis), for the CRP, Fis, and IHF landscapes. Each circle represents visitation data from a different genotype, i.e., the total number of times the genotype was visited across all adaptive walks. We computed regulation strength as a continuous weighted average of fluorescence across sorting bins based on sequencing read distributions (Methods), and incorporated experimental uncertainty as described in **Supplementary Methods 10**. Point density is indicated by color, with warmer colors denoting higher densities of genotypes. The red line shows the best-fit linear model. Spearman correlation coefficients (R) together with p-values for a test of the null hypothesis that genotype visitation rankings are uncorrelated between noisy and noise-free landscapes. The strong correlations observed for all three landscapes indicate that genotype visitation frequencies—and thus landscape-level evolutionary accessibility patterns—are robust to realistic levels of experimental uncertainty. R^2 : coefficient of determination. The panels show correlation plots for **a**. CRP, **b**. Fis, **c**. IHF.

Data availability

Sequencing data has been deposited in the NCBI database under the BioProject accession code: PRJNA1162449. The data generated in this study have been deposited in the Zenodo public repository and are accessible via the following DOI: 10.5281/zenodo.13838265. All code used for processing data and plotting, as well as the final processed data, plasmid sequences, and primer sequences are available in our GitHub repository. The computer code generated in this study has been deposited in the Zenodo public repository and is accessible via the following DOI: 10.5281/zenodo.13838265.

Acknowledgements

We gratefully acknowledge financial support from the Swiss National Science Foundation grant 310030_208174. We also extend our thanks to the UZH University Priority Research Program in Evolutionary Biology, the UZH flow cytometry facility, and the Functional Genomics Center Zurich for their technical support. We are especially thankful to Andrei Papkou for his guidance with computational analysis and for engaging in theoretical discussions.

Additional information

Author contributions

C.A.W. and A.W. conceived the study and designed experiments. C.A.W. carried out experiments. C.A.W. and A.W. analyzed data. C.A.W. wrote computer code to carry out bioinformatic work and analysis. C.A.W. generated figures. L.G. wrote computer code to carry out simulations. C.A.W. and A.W. wrote the paper, which was edited by all authors.

Funding

Funder	Grant reference number	Author
Swiss National Science Foundation	310030_208174	Cauã Antunes Antunes Westmann Leander Goldbach

Author ORCID iDs

Cauã Antunes Westmann: <https://orcid.org/0000-0003-1572-2279>

References

1. **Browning D. F.**, Busby S. J. W (2016) Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol* **14**:638-650 <https://doi.org/10.1038/nrmicro.2016.103> | PubMed
2. **Cases I.**, De Lorenzo V., Ouzounis C. A (2003) Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology* **11**:248-253 [https://doi.org/10.1016/S0966-842X\(03\)00103-3](https://doi.org/10.1016/S0966-842X(03)00103-3) | PubMed
3. **Seshasayee A. S.**, Bertone P., Fraser G. M., Luscombe N. M (2006) Transcriptional regulatory networks in bacteria: from input signals to output responses. *Curr Opin Microbiol* **9**:511-519 <https://doi.org/10.1016/j.mib.2006.08.007> | PubMed
4. **Barnard A.**, Wolfe A., Busby S (2004) Regulation at complex bacterial promoters: How bacteria use different promoter organizations to produce different regulatory outcomes. *Curr Opin Microbiol* **7**:102-108 <https://doi.org/10.1016/j.mib.2004.02.011> | PubMed
5. **Browning D. F. D. F.**, Busby S. J. W. S (2004) The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**:57-65 <https://doi.org/10.1038/nrmicro787> | PubMed

6. Aguilar-Rodríguez J., Payne J. L., Wagner A (2017) A thousand empirical adaptive landscapes and their navigability. *Nat Ecol Evol* **1**:0045 <https://doi.org/10.1038/s41559-016-0045> | PubMed
7. Sharon E., et al. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**:521-530 <https://doi.org/10.1038/nbt.2205> | PubMed
8. Shultzaberger R. K., Malashock D. S., Kirsch J. F., Eisen M. B (2010) The Fitness Landscapes of cis-Acting Binding Sites in Different Promoter and Environmental Contexts. *PLoS Genet* **6**:e1001042 <https://doi.org/10.1371/journal.pgen.1001042> | PubMed
9. Haldane A., Manhart M., Morozov A. V (2014) Biophysical Fitness Landscapes for Transcription Factor Binding Sites. *PLoS Comput Biol* **10**:36-38 <https://doi.org/10.1371/journal.pcbi.1003683> | PubMed
10. Shen-Orr S. S., Milo R., Mangan S., Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**:64-68 <https://doi.org/10.1038/ng881> | PubMed
11. Martínez-Antonio A., et al. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* **6**:482-489 <https://doi.org/10.1016/j.mib.2003.09.002> | PubMed
12. Lozada-Chávez I., Angarica V. E., Collado-Vides J., Contreras-Moreira B (2008) The Role of DNA-binding Specificity in the Evolution of Bacterial Regulatory Networks. *J Mol Biol* **379**:627-643 <https://doi.org/10.1016/j.jmb.2008.04.008> | PubMed
13. Kurafeiski J. D., Pinto P., Bornberg-Bauer E (2019) Evolutionary potential of cis-regulatory mutations to cause rapid changes in transcription factor binding. *Genome Biol Evol* **11**:406-414 <https://doi.org/10.1093/gbe/evy269> | PubMed
14. Browning D. F., Butala M., Busby S. J. W (2019) Bacterial Transcription Factors: Regulation by Pick “N” Mix. *J Mol Biol* **431**:4067-4077 <https://doi.org/10.1016/j.jmb.2019.04.011> | PubMed
15. Visweswariah S. S., Busby S. J. W (2015) Evolution of bacterial transcription factors: How proteins take on new tasks, but do not always stop doing the old ones. *Trends Microbiol* **23**:463-467 <https://doi.org/10.1016/j.tim.2015.04.009> | PubMed
16. Vaishnav E. D., et al. (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**:455-463 <https://doi.org/10.1038/s41586-022-04506-6> | PubMed
17. Schweizer G., Wagner A (2021) Both Binding Strength and Evolutionary Accessibility Affect the Population Frequency of Transcription Factor Binding Sequences in Arabidopsis thaliana. *Genome Biol Evol* **13** <https://doi.org/10.1093/gbe/evab273> | PubMed
18. Westmann C. A., Goldbach L., Wagner A (2024) The highly rugged yet navigable regulatory landscape of the bacterial transcription factor TetR. *Nat Commun* **15** <https://doi.org/10.1038/s41467-024-54723-y> | PubMed
19. Westmann C. A., Goldbach L., Wagner A (2024) Entangled adaptive landscapes facilitate the evolution of gene regulation by exaptation. *bioRxiv* <https://doi.org/10.1101/2024.11.10.620926>
20. Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution.
21. Wright S (1931) Evolution in Mendelian Populations. *Genetics* **16** <https://doi.org/10.1093/genetics/16.2.97> | PubMed
22. Blanco C., Janzen E., Pressman A., Saha R., Chen I. A (2019) Molecular Fitness Landscapes from High-Coverage Sequence Profiling. *Annu Rev Biophys* **48**:1-18 <https://doi.org/10.1146/annurev-biophys-052118-115333> | PubMed
23. Yi X., Dean A. M (2019) Adaptive Landscapes in the Age of Synthetic Biology. *Mol Biol Evol* **36**:890-907 <https://doi.org/10.1093/molbev/msz004> | PubMed
24. Bank C. (2022) Epistasis and Adaptation on Fitness Landscapes. *Annual Review Ecology, Evolution, and Systematics* **53**:457-479 <https://doi.org/10.1146/annurev-ecolsys-102320-112153>
25. Fragata I., Blanckaert A., Dias Louro M. A., Liberles D. A., Bank C (2019) Evolution in the light of fitness landscape theory. *Trends Ecol Evol* **34**:69-82 <https://doi.org/10.1016/j.tree.2018.10.009> | PubMed

26. Otwinowski J., Nemenman I (2013) Genotype to Phenotype Mapping and the Fitness Landscape of the E. coli lac Promoter. *PLoS One* **8**:e61570 <https://doi.org/10.1371/journal.pone.0061570> | PubMed
27. Chattopadhyay G., Papkou A., Wagner A (2025) The fitness landscape of the E.coli lac operator is highly rugged in two different environments. *bioRxiv* <https://doi.org/10.1101/2025.07.23.666252>
28. Herrera-Álvarez S., Patton J. E. J., Thornton J. W (2025) The structure of an ancient genotype–phenotype map shaped the functional evolution of a protein family. *Nat Ecol Evol* **9**:1656–1669 <https://doi.org/10.1038/s41559-025-02777-6> | PubMed
29. Weeks R., Ostermeier M (2023) Fitness and Functional Landscapes of the E. coli RNase III Gene rnc. *Mol Biol Evol* **40** <https://doi.org/10.1093/molbev/msad047> | PubMed
30. Steinberg B., Ostermeier M (2016) Environmental changes bridge evolutionary valleys. *Sci Adv* **2**:e1500921 <https://doi.org/10.1126/sciadv.1500921> | PubMed
31. Papkou A., Garcia-Pastor L., Escudero J. A., Wagner A (2023) A rugged yet easily navigable fitness landscape. *Science* **382**:eadh3860 <https://doi.org/10.1126/science.adh3860> | PubMed
32. Sarkisyan K. S., et al. (2016) Local fitness landscape of the green fluorescent protein. *Nature* **533**:397–401 <https://doi.org/10.1038/nature17995> | PubMed
33. Schuster P (2002) A testable genotype-phenotype map: modeling evolution of RNA molecules. *Biological Evolution and Statistical Physics* 55–81 https://doi.org/10.1007/3-540-45692-9_4
34. García-Galindo P., Ahnert S. E., Martin N. S (2023) The non-deterministic genotype–phenotype map of RNA secondary structure. *J R Soc Interface* **20** <https://doi.org/10.1098/rsif.2023.0132> | PubMed
35. Bendixsen D. P., Collet J., Østman B., Hayden E. J (2019) Genotype network intersections promote evolutionary innovation. *PLoS Biol* **17**:e3000300 <https://doi.org/10.1371/journal.pbio.3000300> | PubMed
36. de Visser J. A. G. M., Cooper T. F., Elena S. F (2011) The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences* **278**:3617–3624 <https://doi.org/10.1098/rspb.2011.1537> | PubMed
37. Wray G. A (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**:206–216 <https://doi.org/10.1038/nrg2063> | PubMed
38. Signor S. A., Nuzhdin S. V (2018) The Evolution of Gene Expression in cis and trans. *Trends in Genetics* **34**:532–544 <https://doi.org/10.1016/j.tig.2018.03.007> | PubMed
39. Wittkopp P. J., Haerum B. K., Clark A. G (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **430**:85–88 <https://doi.org/10.1038/nature02698> | PubMed
40. Yona A. H., Alm E. J., Gore J (2018) Random sequences rapidly evolve into de novo promoters. *Nat Commun* **9** <https://doi.org/10.1038/s41467-018-04026-w> | PubMed
41. Fuqua T., Wagner A (2023) Mobile DNA is replete with hotspots for the de novo emergence of gene regulation. *bioRxiv* <https://doi.org/10.1101/2023.10.22.563463>
42. McAdams H. H., Srinivasan B., Arkin A. P (2004) The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics* **5**:169–178 <https://doi.org/10.1038/nrg1292> | PubMed
43. Dorman C. J., Bhriain N. N., Dorman M. J (2018) The Evolution of Gene Regulatory Mechanisms in Bacteria. In: Rampelotto P (Ed). *Molecular Mechanisms of Microbial Evolution* pp. 125–152 https://doi.org/10.1007/978-3-319-69078-0_6
44. Kinney J. B., McCandlish D. M (2019) Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annu Rev Genomics Hum Genet* **20** <https://doi.org/10.1146/annurev-genom-083118-014845> | PubMed
45. de Visser J. A. G. M., Elena S. F., Fragata I., Matuszewski S (2018) The utility of fitness landscapes and big data for predicting evolution. *Heredity* **121**:401–405 <https://doi.org/10.1038/s41437-018-0128-4> | PubMed
46. Draghi J. A., Ogbunugafor C. B (2023) Exploring the expanse between theoretical questions and experimental approaches in the modern study of evolvability. *J Exp Zool B Mol Dev Evol* **340**:8–17 <https://doi.org/10.1002/jez.b.23134> | PubMed

47. Louis A. A (2016) Contingency, convergence and hyper-astronomical numbers in biological evolution. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **58**:107-116 <https://doi.org/10.1016/j.shpsc.2015.12.014> | PubMed
48. D'Haeseleer P (2006) What are DNA sequence motifs?. *Nature Biotechnology* **24**:423-425
49. Stormo G. D. (2000) DNA binding sites: Representation and discovery. *Bioinformatics* **16**:6-23 <https://doi.org/10.1093/bioinformatics/16.1.16> | PubMed
50. Stormo G. D., Zhao Y (2010) Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11**:751-760 <https://doi.org/10.1038/nrg2845> | PubMed
51. Zhou T., et al. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* **112**:4654-4659 <https://doi.org/10.1073/pnas.1422023112> | PubMed
52. Mathelier A., et al. (2016) DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst* <https://doi.org/10.1016/j.cels.2016.07.001> | PubMed
53. Ibarra I. L., et al. (2020) Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nature Communications* **11**:1-16 <https://doi.org/10.1038/s41467-019-13888-7> | PubMed
54. Bintu L., et al. (2005) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15**:116-124 <https://doi.org/10.1016/j.gde.2005.02.007> | PubMed
55. Jones D. L., Brewster R. C., Phillips R (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346**:1533-1536 <https://doi.org/10.1126/science.1255301> | PubMed
56. Scholz S. A., et al. (2019) High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. *Cell Syst* <https://doi.org/10.1016/j.cels.2019.02.004> | PubMed
57. Kim D., et al. (2018) Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. *Nucleic Acids Res* <https://doi.org/10.1093/nar/gky069> | PubMed
58. Borirak O., et al. (2015) Time-series analysis of the transcriptome and proteome of Escherichia coli upon glucose repression. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1854**:1269-1279 <https://doi.org/10.1016/j.bbapap.2015.05.017> | PubMed
59. Pal A., Iyer M. S., Srinivasan S., Seshasayee A. S. N., Venkatesh K. V (2022) Global pleiotropic effects in adaptively evolved Escherichia coli lacking CRP reveal molecular mechanisms that define the growth physiology. *Open Biol* **12** <https://doi.org/10.1098/rsob.210206> | PubMed
60. Khankal R., Chin J. W., Ghosh D., Cirino P. C (2009) Transcriptional effects of CRP* expression in Escherichia coli. *J Biol Eng* **3**:1-14 <https://doi.org/10.1186/1754-1611-3-13> | PubMed
61. Nowak-Lovato K., et al. (2013) Binding of Nucleoid-Associated Protein Fis to DNA Is Regulated by DNA Breathing Dynamics. *PLoS Comput Biol* **9** <https://doi.org/10.1371/journal.pcbi.1002881> | PubMed
62. Kahramanoglou C., et al. (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res* **39**:2073-2091 <https://doi.org/10.1093/nar/gkq934> | PubMed
63. Gawade P., Gunjal G., Sharma A., Ghosh P (2020) Reconstruction of transcriptional regulatory networks of Fis and H-NS in Escherichia coli from genome-wide data analysis. *Genomics* **112**:1264-1272 <https://doi.org/10.1016/j.ygeno.2019.07.013> | PubMed
64. Prieto A. I., et al. (2012) Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Res* **40**:3524-3537 <https://doi.org/10.1093/nar/gkr1236> | PubMed
65. Arfin S. M., et al. (2000) Global gene expression profiling in Escherichia coli K12. The effects of integration host factor. *J Biol Chem* **275**:29672-29684 <https://doi.org/10.1074/jbc.m002247200> | PubMed

66. Taylor L. R., Provine W. B (2006) Sewall Wright and Evolutionary Biology. *J Anim Ecol* <https://doi.org/10.2307/5082>
67. Lynch M., et al. (2016) Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**:704-714 <https://doi.org/10.1038/nrg.2016.104> | PubMed
68. Iwasa Y., Michor F., Nowak M. A (2004) Stochastic tunnels in evolutionary dynamics. *Genetics* **166**:1571-1579 <https://doi.org/10.1534/genetics.166.3.1571> | PubMed
69. Taylor C. F., Higgs P. G (2000) A population genetics model for multiple quantitative traits exhibiting pleiotropy and epistasis. *J Theor Biol* **203**:419-437 <https://doi.org/10.1006/jtbi.2000.1094> | PubMed
70. Salgado H., et al. (2024) RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res* **52**:D255-D264
71. Peterman N., Levine E (2016) Sort-seq under the hood: Implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**:1-17 <https://doi.org/10.1186/s12864-016-2533-5> | PubMed
72. de Boer C. G., et al. (2019) Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology* **38**:56-65 <https://doi.org/10.1038/s41587-019-0315-8> | PubMed
73. Feng H., et al. (2023) Deep-learning-assisted Sort-Seq enables high-throughput profiling of gene expression characteristics with high precision. *Sci Adv* **9**:eadg5296 <https://doi.org/10.1126/sciadv.adg5296> | PubMed
74. Garcia H. G., Lee H. J., Boedicker J. Q., Phillips R (2011) Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. *Biophys J* **101** <https://doi.org/10.1016/j.bpj.2011.06.026> | PubMed
75. Friedlander T., Prizak R., Barton N. H., Tkačik G (2017) Evolution of new regulatory functions on biophysically realistic fitness landscapes. *Nat Commun* **8** <https://doi.org/10.1038/s41467-017-00238-8> | PubMed
76. Le D. D., et al. (2018) Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proceedings of the National Academy of Sciences* **115** <https://doi.org/10.1073/pnas.1715888115> | PubMed
77. Maerkl S. J. S. J., Quake S. R (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315** <https://doi.org/10.1126/science.1131007> | PubMed
78. Poelwijk F. J., Kiviet D. J., Weinreich D. M., Tans S. J (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* <https://doi.org/10.1038/nature05451> | PubMed
79. Nghe P., et al. (2020) Predicting Evolution Using Regulatory Architecture. *Annu Rev Biophys* **49**:181-197 <https://doi.org/10.1146/annurev-biophys-070317-032939> | PubMed
80. Struhl K (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**:1-4 [https://doi.org/10.1016/s0092-8674\(00\)80599-1](https://doi.org/10.1016/s0092-8674(00)80599-1) | PubMed
81. Barnes S. L., Belliveau N. M., Ireland W. T., Kinney J. B., Phillips R (2019) Mapping DNA sequence to transcription factor binding energy in vivo. *PLoS Comput Biol* **15**:e1006226 <https://doi.org/10.1371/journal.pcbi.1006226> | PubMed
82. Kolb A., Spassky A., Chapon C., Blazy B., Buc H (1983) On the different binding affinities of CRP at the lac, gal and malt promoter regions. *Nucleic Acids Res* <https://doi.org/10.1093/nar/11.22.7833> | PubMed
83. Shao Y., Feldman-Cohen L. S., Osuna R (2008) Functional Characterization of the Escherichia coli Fis-DNA Binding Sequence. *J Mol Biol* **376**:771-785 <https://doi.org/10.1016/j.jmb.2007.11.101> | PubMed
84. Huo Y. X., et al. (2009) IHF-binding sites inhibit DNA loop formation and transcription initiation. *Nucleic Acids Res* **37**:3878-3886 <https://doi.org/10.1093/nar/gkp258> | PubMed
85. Gunasekera A., Ebricht Y. W., Ebricht R. H (1992) DNA sequence determinants for binding of the Escherichia coli catabolite gene activator protein. *Journal of Biological Chemistry* **267**:14713-14720 [https://doi.org/10.1016/s0021-9258\(18\)42099-6](https://doi.org/10.1016/s0021-9258(18)42099-6) | PubMed

86. Aiyar S. E., et al. (2002) Architecture of fis-activated transcription complexes at the Escherichia coli rrnB P1 and rrnE P1 promoters. *J Mol Biol* **316**:501-516 <https://doi.org/10.1006/jmbi.2001.5390> | PubMed
87. Ho S (2013) Wesleyan University.
88. Song S., Zhang J (2021) Unbiased inference of the fitness landscape ruggedness from imprecise fitness estimates. *Evolution* **75**:2658-2671 <https://doi.org/10.1111/evo.14363> | PubMed
89. Obolski U., Ram Y., Hadany L (2018) Key issues review: evolution on rugged adaptive landscapes. *Reports on Progress in Physics* **81** <https://doi.org/10.1088/1361-6633/aa94d4> | PubMed
90. Kauffman S., Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**:11-45 [https://doi.org/10.1016/s0022-5193\(87\)80029-2](https://doi.org/10.1016/s0022-5193(87)80029-2) | PubMed
91. Weinberger E (1990) Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol Cybern* **63**:325-336 <https://doi.org/10.1007/bf00202749>
92. Orr H. A (2003) A Minimum on the Mean Number of Steps Taken in Adaptive Walks. *J Theor Biol* **220**:241-247 <https://doi.org/10.1006/jtbi.2003.3161> | PubMed
93. Saona R., Kondrashov F. A., Khudiakova K. A (2022) Relation Between the Number of Peaks and the Number of Reciprocal Sign Epistatic Interactions. *Bull Math Biol* **84** <https://doi.org/10.1007/s11538-022-01029-z> | PubMed
94. Poelwijk F. J., Tănase-Nicola S., Kiviet D. J., Tans S. J (2011) Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol* **272**:141-144 <https://doi.org/10.1016/j.jtbi.2010.12.015> | PubMed
95. Kvitek D. J., Sherlock G (2011) Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet* <https://doi.org/10.1371/journal.pgen.1002056> | PubMed
96. Conrad M (1990) The geometry of evolution. *Biosystems* **24**:61-81 [https://doi.org/10.1016/0303-2647\(90\)90030-5](https://doi.org/10.1016/0303-2647(90)90030-5) | PubMed
97. Krug J., Oros D (2023) Evolutionary accessibility of random and structured fitness landscapes. *arxiv* <https://arxiv.org/abs/2311.17432v2>
98. Kimura M (1962) On the Probability of Fixation of Mutant Genes in a Population. *Genetics* **47**:713-719
99. Bank C., Matuszewski S., Hietpas R. T., Jensen J. D (2016) On the (un)predictability of a large intragenic fitness landscape. *Proc Natl Acad Sci U S A* **113**:14085-14090 <https://doi.org/10.1073/pnas.1612676113> | PubMed
100. Orr H. A (2002) The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56**:1317-1330 <https://doi.org/10.1111/j.0014-3820.2002.tb01446.x> | PubMed
101. Gillespie J. H (1984) Molecular evolution over the mutational landscape. *Evolution* **38**:1116-1129 <https://doi.org/10.1111/j.1558-5646.1984.tb00380.x> | PubMed
102. Lind P. A., Andersson D. I (2008) Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences* **105**:17878-17883 <https://doi.org/10.1073/pnas.0804445105> | PubMed
103. Horton J. S., Taylor T. B (2023) Mutation bias and adaptation in bacteria. *Microbiology* **169** <https://doi.org/10.1099/mic.0.001404> | PubMed
104. Lee H., Popodi E., Tang H., Foster P. L (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences* **109** <https://doi.org/10.1073/pnas.1210309109> | PubMed
105. Melissa M. J., Good B. H., Fisher D. S., Desai M. M (2022) Population genetics of polymorphism and divergence in rapidly evolving populations. *Genetics* **221** <https://doi.org/10.1093/genetics/iyac053> | PubMed
106. Li J., Amado A., Bank C (2023) Rapid adaptation of recombining populations on tunable fitness landscapes. *Mol Ecol* <https://doi.org/10.1111/MEC.16900> | PubMed

107. Stolyarova A. V., et al. (2022) Complex fitness landscape shapes variation in a hyperpolymorphic species. *eLife* **11** <https://doi.org/10.7554/elife.76073> | PubMed
108. Park S.-C., Neidhart J., Krug J (2016) Greedy adaptive walks on a correlated fitness landscape. *J Theor Biol* **397**:89-102 <https://doi.org/10.1016/j.jtbi.2016.02.035> | PubMed
109. Kauffman S., Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**:11-45 [https://doi.org/10.1016/s0022-5193\(87\)80029-2](https://doi.org/10.1016/s0022-5193(87)80029-2) | PubMed
110. Orr H. A (2003) A Minimum on the Mean Number of Steps Taken in Adaptive Walks. *J Theor Biol* **220**:241-247 <https://doi.org/10.1006/jtbi.2003.3161> | PubMed
111. Blount Z. D., Lenski R. E., Losos J. B (2018) Contingency and determinism in evolution: Replaying life's tape. *Science* **362** <https://doi.org/10.1126/science.aam5979> | PubMed
112. Nonoyama T., Chiba S (2019) Phenotypic determinism and contingency in the evolution of hypothetical tree-like organisms. *PLoS One* **14** <https://doi.org/10.1371/journal.pone.0211671> | PubMed
113. Xie V. C., Pu J., Metzger B. P. H., Thornton J. W., Dickinson B. C (2021) Contingency and chance erase necessity in the experimental evolution of ancestral proteins. *eLife* **10**:1-87 <https://doi.org/10.7554/elife.67336> | PubMed
114. Spor A., et al. (2014) Phenotypic and genotypic convergences are influenced by historical contingency and environment in yeast. *Evolution* **68** <https://doi.org/10.1111/evo.12302> | PubMed
115. Lenski R. E (2017) Convergence and Divergence in a Long-Term Experiment with Bacteria. *Am Nat* **190**:S57-S68 <https://doi.org/10.1086/691209> | PubMed
116. Lind P. A., Farr A. D., Rainey P. B (2015) Experimental evolution reveals hidden diversity in evolutionary pathways. *eLife* **2015** <https://doi.org/10.7554/elife.07074> | PubMed
117. Trippe B. L., et al. (2022) Randomized gates eliminate bias in sort-seq assays. *Protein Science* **31**:e4401
118. Gilliot P. A., Gorochowski T. E (2023) Effective design and inference for cell sorting and sequencing based massively parallel reporter assays. *Bioinformatics* **39** <https://doi.org/10.1093/bioinformatics/btad277> | PubMed
119. Evfratov S. A., et al. (2017) Application of sorting and next generation sequencing to study 5'-UTR influence on translation efficiency in Escherichia coli. *Nucleic Acids Res* **45**:3487-3502 <https://doi.org/10.1093/nar/gkw1141> | PubMed
120. Cuperus J. T., et al. (2017) Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res* **163**:1-10 <https://doi.org/10.1101/gr.224964.117> | PubMed
121. Belliveau N. M., et al. (2018) A Systematic and Scalable Approach for Dissecting the Molecular Mechanisms of Transcriptional Regulation in Bacteria. *Biophys J* **114** <https://doi.org/10.1016/j.bpj.2017.11.849>
122. Kinney J. B., Murugan A., Callan C. G., Cox E. C (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* **107**:9158-9163 <https://doi.org/10.1073/pnas.1004290107> | PubMed
123. Lagator M., et al. (2022) Predicting bacterial promoter function and evolution from random sequences. *eLife* **11** <https://doi.org/10.7554/elife.64543> | PubMed
124. Urtecho G., Tripp A. D., Insigne K. D., Kim H., Kosuri S (2019) Systematic Dissection of Sequence Elements Controlling σ 70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **58**:1539-1551 <https://doi.org/10.1021/acs.biochem.7b01069> | PubMed
125. Aguilar-Rodríguez J., Payne J. L (2021) Robustness and Evolvability in Transcriptional Regulation. *Evolutionary Systems Biology* 197-219 https://doi.org/10.1007/978-3-030-71737-7_9
126. Lagator M., Iglar C., Moreno A. B., Guet C. C., Bollback J. P (2016) Epistatic interactions in the arabinose cis-regulatory element. *Mol Biol Evol* **33**:761-769 <https://doi.org/10.1093/molbev/msv269> | PubMed

127. Lagator M., Paixão T., Barton N. H., Bollback J. P., Guet C. C (2017) On the mechanistic nature of epistasis in a canonical cis-regulatory element. *eLife* **6**:e25192 <https://doi.org/10.7554/elife.25192> | PubMed
128. Stewart A. J., Hannenhalli S., Plotkin J. B., Hannenhalli S., Plotkin J. B (2012) Why Transcription Factor Binding Sites Are Ten Nucleotides Long. *Genetics* **192**:973-985 <https://doi.org/10.1534/genetics.112.143370> | PubMed
129. Babu M. M., Teichmann S. A (2003) Functional determinants of transcription factors in Escherichia coli: Protein families and binding sites. *Trends in Genetics* **19**:75-79 [https://doi.org/10.1016/s0168-9525\(02\)00039-2](https://doi.org/10.1016/s0168-9525(02)00039-2) | PubMed
130. Huffman J. L., Brennan R. G (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Curr Opin Struct Biol* **12**:98-106 [https://doi.org/10.1016/s0959-440x\(02\)00295-6](https://doi.org/10.1016/s0959-440x(02)00295-6) | PubMed
131. Perez-Rueda E., et al. (2018) Abundance, diversity and domain architecture variability in prokaryotic DNA-binding transcription factors. *PLoS One* **13** <https://doi.org/10.1371/journal.pone.0195332> | PubMed
132. Oren Y., et al. (2014) Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences* **111**:16112-16117 <https://doi.org/10.1073/pnas.1413272111> | PubMed
133. Price M. N., Dehal P. S., Arkin A. P (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome Biol* **9** <https://doi.org/10.1186/gb-2008-9-1-r4> | PubMed
134. Weinreich D. M., Watson R. A., Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* <https://doi.org/10.1111/j.0014-3820.2005.tb01768.x> | PubMed
135. Gerland U., Moroz J. D., Hwa T (2002) Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences* **99**:12015-12020 <https://doi.org/10.1073/pnas.192693599> | PubMed
136. Zhao Y., Ruan S., Pandey M., Stormo G. D (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* <https://doi.org/10.1534/genetics.112.138685> | PubMed
137. O'Flanagan R. A., Paillard G., Lavery R., Sengupta A. M (2005) Non-additivity in protein-DNA binding. *Bioinformatics* **21**:2254-2263
138. Payne J. L., Wagner A (2014) The Robustness and Evolvability of Transcription Factor Binding Sites. *Science* **343** <https://doi.org/10.1126/science.1249046> | PubMed
139. Blount Z. D., Borland C. Z., Lenski R. E (2008) Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc Natl Acad Sci U S A* **105**:7899-7906 <https://doi.org/10.1073/pnas.0803151105> | PubMed
140. Palmer A. C., et al. (2015) Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nature Communications* **6**:1-8 <https://doi.org/10.1038/ncomms8385> | PubMed
141. Vermeij G. J (2006) Historical contingency and the purported uniqueness of evolutionary innovations. *Proceedings of the National Academy of Sciences* **103**:1804-1809 <https://doi.org/10.1073/pnas.0508724103> | PubMed
142. Lässig M., Mustonen V., Walczak A. M (2017) Predicting evolution. *Nat Ecol Evol* **1**:1-9 <https://doi.org/10.1038/s41559-017-0077> | PubMed
143. De Visser J. A. G. M., Krug J (2014) Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* **15**:480-490 <https://doi.org/10.1038/nrg3744> | PubMed
144. Heyde S. A. H., Frensdorff P. O., Lauritsen I., Nørholm M. H. H (2021) Restoring Global Gene Regulation through Experimental Evolution Uncovers a NAP (Nucleoid-Associated Protein)-Like Behavior of Crp/Cap. *mBio* **12** <https://doi.org/10.1128/mbio.02028-21> | PubMed

145. Monteiro L. M. O., Sanches-Medeiros A., Westmann C. A., Silva-Rocha R (2020) Unraveling the Complex Interplay of Fis and IHF Through Synthetic Promoter Engineering. *Front Bioeng Biotechnol* **8** <https://doi.org/10.3389/fbioe.2020.00510> | PubMed
146. Dillon S. C., Dorman C. J (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat Rev Microbiol* **8**:185-95 <https://doi.org/10.1038/nrmicro2261> | PubMed
147. Dorman C. J (2013) Genome architecture and global gene regulation in bacteria: Making progress towards a unified model?. *Nat Rev Microbiol* **11**:349-355 <https://doi.org/10.1038/nrmicro3007> | PubMed
148. Dorman C. J., Schumacher M. A., Bush M. J., Brennan R. G., Buttner M. J (2020) When is a transcription factor a NAP?. *Curr Opin Microbiol* **55** <https://doi.org/10.1016/j.mib.2020.01.019> | PubMed
149. Chen Y. J., et al. (2020) Quantifying molecular bias in DNA data storage. *Nature Communications* **11**:1-9 <https://doi.org/10.1038/s41467-020-16958-3> | PubMed
150. Kebschull J. M., Zador A. M (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* **43**:e143-e143 <https://doi.org/10.1093/nar/gkv717> | PubMed
151. Wong T. S., Zhurina D., Schwaneberg U (2006) The Diversity Challenge in Directed Protein Evolution. *Comb Chem High Throughput Screen* **9**:271-288 <https://doi.org/10.2174/138620706776843192> | PubMed
152. Aird D., et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**:1-14 <https://doi.org/10.1186/gb-2011-12-2-r18> | PubMed
153. Gilliot P.-A., Gorochowski T. E (2022) Effective design and inference for cell sorting and sequencing based massively parallel reporter assays. *bioRxiv* <https://doi.org/10.1101/2022.11.07.515414>
154. Wagner A (2022) AI predicts the effectiveness and evolution of gene promoter sequences. *Nature* **603**:399-400 <https://doi.org/10.1038/d41586-022-00384-0> | PubMed
155. Boer C. de, Sadeh R., Friedman N., Regev A (2018) Deciphering cis-regulatory logic with 100 million random promoters. *bioRxiv* <https://doi.org/10.1101/224907>
156. De Boer C. G., Ray J. P., Hacohen N., Regev A (2020) MAUDE: Inferring expression changes in sorting-based CRISPR screens. *Genome Biol* **21**:1-16 <https://doi.org/10.1186/s13059-020-02046-8> | PubMed
157. Wu N. C., Dai L., Olson C. A., Lloyd-Smith J. O., Sun R (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5** <https://doi.org/10.7554/elife.16965> | PubMed
158. Diaz-Colunga J., et al. (2023) Global epistasis on fitness landscapes. *Philosophical Transactions of the Royal Society B* **378** <https://doi.org/10.1098/rstb.2022.0053> | PubMed
159. MacLean R. C., Perron G. G., Gardner A (2010) Diminishing Returns From Beneficial Mutations and Pervasive Epistasis Shape the Fitness Landscape for Rifampicin Resistance in *Pseudomonas aeruginosa*. *Genetics* **186**:1345-1354 <https://doi.org/10.1534/genetics.110.123083> | PubMed
160. Rydenfelt M., Garcia H. G., Cox R. S., Phillips R (2014) The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS One* **9**:1-31 <https://doi.org/10.1371/journal.pone.0114347> | PubMed
161. Ishihama A (2010) Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol Rev* **34**:628-645 <https://doi.org/10.1111/j.1574-6976.2010.00227.x> | PubMed
162. Cases I., de Lorenzo V (2005) Promoters in the environment: Transcriptional regulation in its natural context. *Nature Reviews Microbiology* **3**:105-118 <https://doi.org/10.1038/nrmicro1084> | PubMed
163. Shepherd M. J., Reynolds M., Pierce A. P., Rice A. M., Taylor T. B (2023) Transcription factor expression levels and environmental signals constrain transcription factor innovation. *Microbiology* **169** <https://doi.org/10.1099/mic.0.001378> | PubMed
164. González Pérez A., Espinosa Angarica V., Collado-Vides J., Vasconcelos A. T. R (2009) From sequence to dynamics: The effects of transcription factor and polymerase concentration changes on activated and repressed promoters. *BMC Mol Biol* **10** <https://doi.org/10.1186/1471-2199-10-92> | PubMed

165. Azam T. A., Iwata A., Nishimura A., Ueda S., Ishihama A (1999) Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J Bacteriol* <https://doi.org/10.1128/jb.181.20.6361-6370.1999> | PubMed
166. Weinert F. M., et al. (2014) The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Biophys J* <https://doi.org/10.1016/j.bpj.2013.11.4483>
167. Lagomarsino M. C., Espéli O., Junier I (2015) From structure to function of bacterial chromosomes: Evolutionary perspectives and ideas for new experiments. *FEBS Letters* <https://doi.org/10.1016/j.febslet.2015.07.002> | PubMed
168. Dorman C. J., Dorman M. J (2016) DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. *Biophys Rev* **8**:209-220 <https://doi.org/10.1007/s12551-016-0205-y> | PubMed
169. Houdaigui B. El, et al. (2019) Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. *Nucleic Acids Res* **47**:5648-5657 <https://doi.org/10.1093/nar/gky069>
170. Sobetzko P., Travers A., Muskhelishvili G (2011) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proceedings of the National Academy of Sciences* **109**:E42-E50 <https://doi.org/10.1073/pnas.1108229109> | PubMed
171. Meyer S., Reverchon S., Nasser W., Muskhelishvili G (2018) Chromosomal organization of transcription: in a nutshell. *Curr Genet* **64**:555-565 <https://doi.org/10.1007/s00294-017-0785-5> | PubMed
172. Sánchez-Romero M. A., Cota I., Casadesús J (2015) DNA methylation in bacteria: from the methyl group to the methylome. *Curr Opin Microbiol* **25**:9-16 <https://doi.org/10.1016/j.mib.2015.03.004> | PubMed
173. Crocker J., Preger-Ben Noon E., Stern D. L (2016) The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. In: Wassarman P. M (Ed). *Current Topics in Developmental Biology* **117** <https://doi.org/10.1016/bs.ctdb.2015.11.018> | PubMed
174. Srivastava M., Payne J. L (2022) On the incongruence of genotype-phenotype and fitness landscapes. *PLoS Comput Biol* **18**:e1010524 <https://doi.org/10.1371/journal.pcbi.1010524> | PubMed
175. Rest J. S., et al. (2013) Nonlinear fitness consequences of variation in expression level of a eukaryotic gene. *Mol Biol Evol* **30**:448-456 <https://doi.org/10.1093/molbev/mss248> | PubMed
176. Dubeau F., Toubiana W., Wittkopp P. J (2017) Fitness effects of cis-regulatory variants in the saccharomyces cerevisiae TDH3 promoter. *Mol Biol Evol* **34** <https://doi.org/10.1093/molbev/msx224> | PubMed
177. Bergen A. C., Olsen G. M., Fay J. C (2016) Divergent MLS1 Promoters Lie on a Fitness Plateau for Gene Expression. *Mol Biol Evol* **33** <https://doi.org/10.1093/molbev/msw010> | PubMed
178. Ornelas M. Y., Cournoyer J. E., Bram S., Mehta A. P (2023) Evolution and synthetic biology. *Curr Opin Microbiol* **76**:102394 <https://doi.org/10.1016/j.mib.2023.102394> | PubMed
179. Baba T., et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* **2** <https://doi.org/10.1038/msb4100050> | PubMed
180. Elowitz M. B., Levine A. J., Siggia E. D., Swain P. S (2007) Stochastic Gene Expression in a Single Cell. *Science* **297**:1183-6 <https://doi.org/10.1126/science.1070919> | PubMed
181. Sharon E., et al. (2014) Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res* **24**:1698-1706 <https://doi.org/10.1101/gr.168773.113> | PubMed
182. Rudge T. J., et al. (2016) Characterization of Intrinsic Properties of Promoters. *ACS Synth Biol* **5**:89-98 <https://doi.org/10.1021/acssynbio.5b00116> | PubMed
183. Gibson D. G., et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**:343-5 <https://doi.org/10.1038/nmeth.1318> | PubMed
184. Warren D. J (2011) Preparation of highly efficient electrocompetent Escherichia coli using glycerol/mannitol density step centrifugation. *Anal Biochem* **413**:206-207 <https://doi.org/10.1016/j.ab.2011.02.036> | PubMed

185. Jahn M., Vorpahl C., Hübschmann T., Harms H., Müller S (2016) Copy number variability of expression plasmids determined by cell sorting and droplet digital PCR. *Microb Cell Fact* **15** <https://doi.org/10.1186/s12934-016-0610-8> | PubMed
186. Lutz R., Bujard H (1997) Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res* **25**:1203-10 <https://doi.org/10.1093/nar/25.6.1203> | PubMed
187. Pédelacq J. D., Cabantous S., Tran T., Terwilliger T. C., Waldo G. S (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**:79-88 <https://doi.org/10.1038/nbt1172> | PubMed
188. Lou C., Stanton B., Chen Y. J., Munsy B., Voigt C. A (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol* <https://doi.org/10.1038/nbt.2401> | PubMed
189. Bindels D. S., et al. (2016) mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nature Methods* **14**:53-56 <https://doi.org/10.1038/nmeth.4074> | PubMed
190. Beal J (2017) Biochemical complexity drives log-normal variation in genetic expression. *Engineering Biology* **1**:55-60 <https://doi.org/10.1049/enb.2017.0004>
191. Beal J., et al. (2016) Reproducibility of fluorescent expression from engineered biological constructs in *E. coli*. *PLoS One* **11** <https://doi.org/10.1371/journal.pone.0150182> | PubMed
192. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**:10-12 <https://doi.org/10.14806/ej.17.1.200>
193. Magoč T., Salzberg S. L (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**:2957-2963 <https://doi.org/10.1093/bioinformatics/btr507> | PubMed
194. Hannon G. J (no date) FASTX-Toolkit.
195. Bunn A., Korpela M (2013) A dendrochronology program library in R (dplR). *Dendrochronologia* <https://doi.org/10.1016/j.dendro.2008.01.002>
196. Schneider T. D., Stephens R. M (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**:6097 <https://doi.org/10.1093/nar/18.20.6097> | PubMed
197. Csardi G (2014) The Igraph Software Package for Complex Network Research. <https://www.researchgate.net/publication/221995787>
198. Khalid F., et al. (2016) Genonets server-a web server for the construction, analysis and visualization of genotype networks. *Nucleic Acids Res* **44**:W70-W76 <https://doi.org/10.1093/nar/gkw313> | PubMed
199. Jaccard P (1912) The distribution of the flora in the alpine zone. *New Phytologist* **11**:37-50 <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>
200. Papkou A., Garcia-Pastor L., Escudero J. A., Wagner A (2023) A rugged yet easily navigable fitness landscape of antibiotic resistance. *bioRxiv* <https://doi.org/10.1101/2023.02.27.530293>
201. Kimura M (1983) The Neutral Theory of Molecular Evolution. *Cambridge University Press* <https://doi.org/10.1017/CBO9780511623486>
202. Crow J., Kimura M. (2009) *An Introduction to Population Genetics Theory* **608** The Blackburn Press.
203. Harris C. R., et al. (2020) Array programming with NumPy. *Nature* **585**:357-362 <https://doi.org/10.1038/s41586-020-2649-2> | PubMed
204. Grant S. G., Jessee J., Bloom F. R., Hanahan D (1990) Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. *Proceedings of the National Academy of Sciences* **87**:4645-4649 <https://doi.org/10.1073/pnas.87.12.4645> | PubMed
205. Ebright R. H., Ebright Y. W., Gunasekera A (1989) Consensus DNA site for the *Escherichia coli* catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the *E. coli* lac DNA site. *Nucleic Acids Res* **17**:10295-10305 <https://doi.org/10.1093/nar/17.24.10295> | PubMed

206. Shao Y., Feldman-Cohen L. S., Osuna R (2008) Biochemical Identification of Base and Phosphate Contacts between Fis and a High-Affinity DNA Binding Site. *J Mol Biol* **380**:327-339 <https://doi.org/10.1016/j.jmb.2008.04.075> | PubMed
 207. Kahramanoglou C., et al. (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res* **39**:2073-2091 <https://doi.org/10.1093/nar/gkq934> | PubMed
 208. Kelly J. R., et al. (2009) Measuring the activity of BioBrick promoters using an in vivo reference standard. *J Biol Eng* **3** <https://doi.org/10.1186/1754-1611-3-4> | PubMed
 209. Meyer A. J., Segall-Shapiro T. H., Glassey E., Zhang J., Voigt C. A (2018) Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. *Nature Chemical Biology* **15**:196-204 <https://doi.org/10.1038/s41589-018-0168-3> | PubMed
 210. Carr S. B., Beal J., Densmore D. M (2017) Reducing DNA context dependence in bacterial promoters. *PLoS One* <https://doi.org/10.1371/journal.pone.0176013> | PubMed
 211. Chen Y. J., et al. (2013) Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods* <https://doi.org/10.1038/nmeth.2515> | PubMed
 212. Wickham H (2016) *Ggplot2: Elegant Graphics for Data Analysis* New York: Springer-Verlag.
- Caua Westmann (2024) Sort-seq data from CRP, Fis and IHF. NCBI BioProject. ID PRJNA1162449 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1162449>

Peer reviews

Reviewer #1 (Public review):

Summary:

For each of three key transcription factor (TF) proteins in *E. coli*, the authors generate a large library of TF binding site (TFBS) sequences on plasmids, such that each TFBS is coupled to the expression of a fluorescence reporter. By sorting the fluorescence of individual cells and sequencing their plasmids to identify each cell's TFBS sequence (sort-seq), they are able to map the landscape of these TFBSs to the gene expression level they regulate. The authors then study the topographical features of these landscapes, especially the number and distribution of local maxima, as well as the statistical properties of evolutionary paths on these landscapes. They find the landscapes to be highly rugged, with about as many local peaks as a random landscape would have, and with those peaks distributed approximately randomly in sequence space. This is quite different from previous work on landscapes for eukaryotic TFBSs, which tend to be rather smooth. The authors find that there are a number of peaks that produce regulation stronger than that of the wild-type sequence for each TF, and that it is not too unlikely to reach one of those "high peaks" from a random starting sequence. Nevertheless, the basins of attraction for different peaks have significant overlap, which means that chance plays a major role in determining which peak a population will evolve to.

Strengths:

(1) The apparent differences in landscape topography between prokaryotic TFBSs and other molecular landscapes is a fascinating discovery to add to the field of genotype-phenotype maps. I am really excited to learn the molecular mechanisms of this in the future.

(2) The experiments and analysis of this paper are very well-executed and, by and large, very thorough. I appreciated the systematic nature of the project, both the large-scale experiments done on three TFs with replicates, and the systematic analysis of the resulting landscapes. This not only makes the paper easy to follow, but also inspires confidence in their results

since there is so much data and so many different ways of analyzing it. It's a great recipe for other studies of genotype-phenotype landscapes to follow.

(3) Considering how technical the project was, I am really impressed at how easy to read I found the paper, and the authors deserve a lot of credit for making it so. They do a great job of building up the experiments and analyses step-by-step, and explaining enough of the basics of the experimental design and essence of each analysis in the main text without getting too complicated with details that can be left to the Methods or SI.

Weaknesses:

(1) Regarding the effect of measurement uncertainties, one way in which they attempt to test their effect is to simulate dynamics on noisy and noise-free versions of the landscape and measure visitation frequencies. While they show that visitation frequencies are highly correlated between these cases, I'd prefer a more direct test of epistasis or navigability (e.g., number of local peaks), since that's how they are characterizing the landscapes, and the connection between that and visitation frequency of individual states is unclear.

(2) I am still a little concerned about the fraction of sequences missing from the data due to filtering, although I appreciate the difficulties in testing the importance of this (requiring additional assumptions) and the authors' good-faith efforts to do their best with the data they have.

<https://doi.org/10.7554/eLife.103774.2.sa2>

Reviewer #2 (Public review):

The authors aim to investigate the ability of evolution to create strong transcription factor binding sites (TFBSs) de novo in *E. coli*. They focus on three global transcriptional regulators: CRP, Fis, and IHF, using a massively parallel reporter assay to evaluate the regulatory effects of over 30,000 TFBS variants. By analyzing the resulting genotype-phenotype landscapes, they explore the ruggedness, accessibility, and evolutionary dynamics of regulatory landscapes, providing insights into the evolutionary feasibility of strong gene regulation. Their experiments show that de novo adaptive evolution of new gene regulation is feasible. It is also subject to a blend of chance, historical contingency, and evolutionary biases that favor some peaks and evolutionary paths.

(1) Strengths of the methods and results:

The authors successfully employed a well-designed sort-seq assay combined with high-throughput sequencing to map regulatory landscapes. The experimental design ensures reliable measurement of regulation strengths. Their system accounts for gene expression noise and normalizes measurements using appropriate controls.

Comprehensive Landscape Mapping:

The study examines ~30,000 TFBS variants per transcription factor, providing statistically robust and thorough maps of the regulatory landscapes for CRP, Fis, and IHF. The landscapes are rigorously analyzed for ruggedness (e.g., number of peaks) and epistasis, revealing parallels with theoretical uncorrelated random landscapes.

Evolutionary Dynamics Simulations:

Through simulations of adaptive walks under varying population dynamics, the authors demonstrate that high peaks in regulatory landscapes are accessible despite ruggedness. They identify key evolutionary phenomena, such as contingency (multiple paths to peaks) and biases toward specific evolutionary outcomes.

Biological Relevance and Novelty:

The author's work is novel in focusing on global regulators, which differ from previously studied local regulators (e.g., *TetR*). They provide compelling evidence that rugged landscapes are navigable, facilitating de novo evolution of regulatory interactions. The comparison of landscapes for CRP, Fis, and IHF underscores shared topographical features, suggesting general principles of global transcriptional regulation in bacteria.

(2) Weaknesses of the methods and results:

Undersampling of Genotype Space:

Approximately 40% of the theoretical TFBS genotype space remains uncharacterized after quality filtering. The authors now discuss this limitation more explicitly and provide analyses suggesting that undersampling does not strongly bias their conclusions at the landscape level. Nevertheless, predictive modeling approaches could further extend these landscapes in future work.

Simplified Regulatory Architecture:

The study considers a minimal system consisting of a single TFBS upstream of a reporter gene. While this simplification allows clean interpretation and high-throughput measurement, natural promoters often involve combinatorial regulation and chromosomal context effects that may alter landscape topography.

Lack of Experimental Evolution Validation:

The evolutionary conclusions are based on simulations rather than direct experimental evolution. The authors provide a reasonable justification for this choice and frame their conclusions at the statistical level rather than for specific trajectories, but experimental validation would be a valuable future extension.

Impact on the Field:

This study advances our understanding of adaptive landscapes in gene regulation and offers a critical step toward deciphering how global regulators evolve de novo binding sites. The findings provide foundational insights for synthetic biology, evolutionary genetics, and systems biology by highlighting the evolutionary accessibility of strong regulation in bacteria.

Utility of Methods and Data:

The sort-seq approach, combined with landscape analysis, provides a robust framework that can be extended to other transcription factors and systems. If made publicly available, the study's data and code would be valuable for researchers modeling transcriptional regulation or studying evolutionary dynamics.

Additional Context:

The study builds on a growing body of work exploring regulatory evolution. For instance, recent studies on local regulators like *TetR* and *AraC* have revealed high ruggedness and epistasis in TFBS landscapes. This study distinguishes itself by focusing on global regulators, which are more complex biologically and more influential in bacterial gene networks. The observed evolutionary contingency aligns with findings in other biological systems, such as protein evolution and RNA folding landscapes, underscoring the generality of these evolutionary principles.

Conclusion:

The authors successfully mapped the genotype-phenotype landscapes for three global regulators and simulated evolutionary dynamics to assess the feasibility of strong TFBS evolution. They convincingly demonstrate that ruggedness and epistasis, while prominent, do not preclude the evolution of strong regulation. Their results support the notion that gene regulation evolves through a blend of chance, contingency, and evolutionary biases.

This paper makes a significant contribution to the understanding of regulatory evolution in bacteria. While minor limitations exist, the authors' methods are robust, and their findings are well-supported. The work will likely be of broad interest to researchers in molecular evolution, synthetic biology, and gene regulation.

<https://doi.org/10.7554/eLife.103774.2.sa1>

Author Response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public review):

Weaknesses:

(1) The main weakness of this paper, in my view, is that it felt disconnected from the larger body of work on fitness and genotype-phenotype landscapes, including previous data on TFBSs in E. coli, genotype-phenotype maps of TFBSs in other systems, protein sequence landscapes (e.g., from mutational scans or combinatorially-complete libraries), and fitness landscapes of genomic mutations (e.g., combinatorially-complete landscapes of antibiotic resistance alleles). I have no doubt the authors are experts in this literature, and they probably cite most of it already given the enormous number of references. But they don't systematically introduce and summarize what was already known from all that work, and how their present study builds on it, in the Abstract and Introduction, which left me wondering for most of the paper why this project was necessary. Eventually, the authors do address most of these points, but not until the end, in the Discussion. Readers who have no familiarity with this literature might read this paper thinking that it's the first paper ever to study topography and evolutionary paths on genotype-phenotype landscapes, which is not true.

There were two points that made this especially confusing for me. First, in order to choose which nucleotides in the binding sites to vary, the authors invoke existing data on the diversity of these sequences (position-weight matrices from RegulonDB). But since those PWMs can imply a genotype-phenotype map themselves, an obvious question I think the authors needed to have answered right away in the Introduction is why it is insufficient for their question. They only make a brief remark much later in the Results that the PWM data is just observed sequence diversity and doesn't directly reflect the regulation strength of every possible TFBS sequence. But that is too subtle in my opinion, and such a critical motivation for their study that it should be a major point in the Introduction.

The second point where the lack of motivation in the Introduction created confusion for me was that they report enormous levels of sign epistasis in their data, to the point where these landscapes look like random uncorrelated landscapes. That was really surprising to me since it contrasts with other empirical landscape data I'm familiar with. It was only in the Discussion that I found some significant explanation of this - namely that this could be a difference between prokaryotic TFBSs, as this paper studies, and the eukaryotic TFBSs that have been the focus of many (almost all?) previous work. If that is

in fact the case - that almost all previous studies have focused on eukaryotic TFBSs or other kinds of landscapes, and this is the first to do a systematic test of prokaryotic TFBS, then that should be a clear point made in the Abstract and Introduction. (I find a comparable statement only in the very last paragraph of the Discussion.) If that's the case, then I would also find that point to be a much stronger, more specific conclusion of this paper to emphasize than the more general result of observing epistasis and contingency (as is currently emphasized in the Abstract), which has been discussed in tons of other papers. This raises all sorts of exciting questions for future studies - why do the landscapes of prokaryotic TFBSs differ so dramatically from almost all the other landscapes we've observed in biology? What does that mean for the evolutionary dynamics of these different systems?

We thank the reviewer for this thoughtful and detailed critique. We agree that the original version of the manuscript did not sufficiently motivate the study early on, nor did it clearly position our work within the broader literature on genotype–phenotype (GP) and fitness landscapes. We also agree that two specific issues, the role of PWMs and the unexpectedly high levels of sign epistasis, were insufficiently explained early on, which could lead to confusion for readers not already familiar with this field.

Positioning within the broader landscape literature

In response, we have substantially revised the Abstract and Introduction to explicitly situate our work within existing empirical studies of GP and fitness landscapes, including TFBS landscapes in bacteria, eukaryotic TFBS genotype–phenotype maps, in vitro TF–DNA binding studies, deep mutational scans of proteins, and combinatorially complete fitness landscapes such as antibiotic resistance alleles (Abstract; Introduction, lines 64–85). We now make clear that our study builds directly on this extensive body of work, rather than introducing the landscape framework itself. For example, we write in the introduction:

“Over the last decade, genotype–phenotype (GP) maps and fitness landscapes have become central tools for understanding how molecular systems evolve under mutation and selection[22–25]. Such maps and landscapes have been experimentally studied for DNA[6,8,18,19,26,27], protein[28–32] and RNA[33–35] molecules, revealing key topographical properties that shape evolutionary outcomes, including epistasis[24,36]—the non-additive effects of multiple mutations on phenotype—landscape ruggedness, reflected in the number and distribution of fitness peaks, and constraints on adaptive evolution.”

At the same time, we clarify what remains rare in the literature: large-scale, *in vivo* genotype–phenotype landscapes for bacterial transcription factor binding sites that are sufficiently dense to support explicit evolutionary analyses. While numerous high-throughput studies have characterized bacterial regulatory elements, these datasets typically do not provide quantitative regulatory phenotypes across large genotype spaces, nor do they analyze evolutionary accessibility. To our knowledge, only one such *in vivo* TFBS landscape had previously been characterized at comparable resolution for a bacterial local regulator (*TetR*). Our work extends this approach to three global regulators, enabling systematic comparisons across prokaryotic systems (Abstract, Introduction, lines 64–85). For example, we write in the introduction:

“For transcription factor binding sites, most pertinent large-scale studies are based on in vitro binding assays, such as protein-binding microarrays (PBMs), and they focus predominantly on eukaryotic transcription factors[6]. While these studies have been instrumental in characterizing transcription factor binding preferences, they typically do not measure regulatory output in a native cellular context. In contrast, comprehensive *in vivo* data for bacterial TFBSs remain extremely rare. To our knowledge, only two high-resolution *in vivo* landscapes have been previously mapped for bacterial regulators, those of the local regulators *TetR*[18] and *LacI*[27]. As a result, it remains unclear whether principles inferred

from protein landscapes, eukaryotic TFBSs, or *in vitro* binding assays generalize to transcriptional regulation in bacteria, particularly for global regulators[11] that integrate multiple physiological signals.”

Why PWMs are insufficient for our question.

We agree with the reviewer that our original explanation of the role of PWMs was too cursory and should have been addressed explicitly in the Introduction. We have now revised the Introduction to clearly explain why PWMs derived from RegulonDB cannot substitute for empirical GP landscapes in our study (Introduction, lines 102–113).

In this passage we now explain that, first, PWMs are inferred from a limited number of naturally occurring binding sites—typically on the order of hundreds of sequences—whose diversity reflects evolutionary history and genomic context rather than systematic exploration of sequence space. As a result, PWMs sample only a small and biased subset of the possible TFBS variants, whereas our libraries probe tens of thousands of sequences in a controlled manner, providing substantially broader and more uniform coverage of genotype space (Introduction, lines 102–113).

Second, PWM scores are not direct measurements of regulatory strength. Instead, they represent probabilistic or heuristic scores that are primarily used for identifying candidate binding sites in genomes. Numerous studies have shown that PWM scores often correlate weakly with *in vivo* binding affinity or regulatory output, where DNA shape, cooperative interactions, and chromosomal context play important roles. As such, PWMs do not provide quantitative genotype–phenotype relationships for regulation strength (Introduction, lines 102–113).

Third, PWMs assume independent and additive contributions of individual nucleotide positions. This assumption excludes epistatic interactions by construction. Because epistasis is central to landscape ruggedness, peak structure, and evolutionary accessibility, PWM-based models are fundamentally unsuited to address the evolutionary questions we study here (Introduction, lines 102–113). We now explicitly state this limitation early in the manuscript, rather than only alluding to it later in the Results.

Sign epistasis and contrast with prior TFBS landscapes.

We also agree with the reviewer that the extensive sign epistasis we observe—approaching levels expected for uncorrelated random landscapes—is surprising in light of much of the existing empirical landscape literature. Importantly, as the reviewer notes, most previous TFBS landscape studies have focused on *in vitro* binding systems or on eukaryotic transcription factors, which tend to exhibit smoother and more additive landscapes.

To address this concern, we have revised the Abstract and Introduction to explicitly frame this contrast as a central result of the study (Abstract; Introduction, lines 151–153, Discussion, lines 652–668). For example, we write in the discussion:

“We showed that the regulatory landscapes of all three TFs are highly rugged and have multiple peaks. The ruggedness of all three landscapes is also supported by the prevalence of epistasis between pairs of TFBS mutations (Supplementary Table S5). A particularly important form of epistasis is sign epistasis[24,93,94], because it can lead to multiple adaptive peaks [24,93,94] (see Supplementary Methods 7.5). Our landscapes contain up to 65% of mutation pairs with sign epistasis, a value that is especially high compared to the almost exclusively additive interactions of mutations in eukaryotic TFs[6,125].”

We now emphasize that prokaryotic TFBS landscapes, particularly for global regulators, appear to be substantially more rugged and epistatic than most previously characterized

TFBS landscapes, and that this difference likely reflects fundamental biological distinctions between regulatory systems.

Revised emphasis and conclusions.

Following the reviewer's suggestion, we have adjusted the emphasis of the manuscript accordingly. Rather than highlighting epistasis and contingency as generic evolutionary phenomena, we now present the extreme ruggedness of prokaryotic TFBS landscapes as a system-specific finding with important implications for the evolution of gene regulation. We explicitly note that this raises new questions for future work—such as why prokaryotic regulatory landscapes differ so markedly from eukaryotic ones, and how these differences shape evolutionary dynamics—which we now highlight in the Introduction and Discussion (Abstract; Introduction, lines 151-153, Discussion, lines 652–668). For example, we write in the discussion:

“... A possible reason for this greater incidence of epistasis lies in the nature of prokaryotic TFBSs. Specifically, prokaryotic TFBSs are at approximately 20bps twice as long as eukaryotic TFBSs[80,128] and exhibit symmetries that reflect the dimeric state of their cognate TFs[129–131]. These factors may increase the likelihood of intramolecular epistasis. Our observations raise important questions for future work, such as why the landscapes of prokaryotic TFBSs differ so dramatically from those of eukaryotic ones. And what do these differences imply for the evolutionary dynamics of gene regulation?”

We believe that these revisions substantially improve the clarity, motivation, and positioning of the manuscript, and directly address the reviewer's concerns by making both the necessity and the novelty of the study clear from the outset.

(2) I am a bit concerned about the lack of uncertainties incorporated into the results. The authors acknowledge several key limitations of their approach, including the discreteness of the sort-seq bins in determining possible values of regulation strength, the existence of a large number of unsampled sequences in their genotype space, as well as measurement noise in the fluorescence readouts and sequencing. While the authors acknowledge the existence of these factors, I do not see much attempt to actually incorporate the effect of these uncertainties into their conclusions, which I suspect may be important. For example, given the bin size for the fluorescence in sort-seq, how confident are they that every sequence that appears to be a peak is actually a peak? Is it possible that many of the peak sequences have regulation strengths above all their neighbors but within the uncertainty of the fluorescence, making it possible that it's not really a peak? Perhaps such issues would average out and not change the statistical nature of their results, which are not about claiming that specific sequences are peaks, just how many peaks there are. Nevertheless, I think the lack of this robustness analysis makes the results less convincing than they otherwise would be.

We thank the reviewer for raising this important concern. We fully agree that uncertainties arising from experimental resolution, measurement noise in fluorescence and sequencing, and incomplete sampling of genotype space should be incorporated explicitly into the analysis. While these limitations were acknowledged qualitatively in the original manuscript, we recognize that a direct, quantitative assessment of their impact on our conclusions is essential to strengthen the robustness of the study.

We first clarify that regulation strength is not discretized in our analysis. For each TFBS, regulation strength is calculated as a continuous weighted average of fluorescence across all sorting bins, based on the sequencing read-count distribution of each sequence across bins. We clarified this information in the main text (Results, lines 201-203). Nevertheless, finite binning resolution and experimental noise introduce uncertainty in these estimates, which could in principle affect the identification of local peaks.

Importantly, our study does not aim to assert that specific TFBS sequences are definitively peaks. Rather, our focus is on landscape-level statistical and topological properties—such as ruggedness, the abundance and distribution of peaks, and the evolutionary accessibility of strong regulation. We therefore centered our new analyses on testing whether these conclusions are robust to experimentally plausible sources of uncertainty, rather than on the identity of individual peaks.

To address the reviewer’s concern, we performed two complementary analyses. The first evaluates whether the observed ruggedness of the landscapes could arise as an artifact of incomplete sampling. It addressed the effects of missing genotypes and the possibility of spurious peak identification due to unsampled neighbors. Sparse sampling can introduce opposing biases: true peaks may be missed, while other genotypes may be falsely classified as peaks because fitter neighbors are absent. As shown for uncorrelated random (House-of-Cards) landscapes (Kauffman & Levin, 1987), these effects can partially cancel.

In this analysis, we constructed a null model by randomly permuting regulation strengths across the mapped genotype network while preserving its topology. The number of peaks in these randomized landscapes is only modestly higher than in the empirical data, indicating that the measured landscapes are close to the maximal ruggedness compatible with the sampled network (Results, lines 308–320).

In addition, we quantified potential sampling bias by analyzing genotype connectivity. Here we defined the relative connectivity of a genotype as the fraction of possible single-mutant neighbors for which we had measured regulation strength. We observed only a very weak correlation between connectivity and regulation strength ($R=-0.1$, -0.1 , 0.01 for the CRP, Fis, and IHF landscapes, Figures S13–S15). Similarly, the relative connectivity of peak genotypes is only weakly correlated with their regulation strength ($R=-0.05$, -0.04 , 0.06 for the CRP, Fis, and IHF landscapes). (Results, lines 321–330), indicating that strongly regulating genotypes are not preferentially oversampled or undersampled (Results, lines 321–330).

The second, and most important, analysis directly addresses the reviewer’s concern that experimental uncertainty could affect peak classification and, consequently, landscape navigability. We explicitly incorporated experimentally measured, genotype-specific noise estimates from biological replicates when comparing fitness values between neighboring genotypes. Using these uncertainty-aware comparisons, we then recomputed adaptive-walk dynamics and genotype visitation frequencies on the resulting noisy landscapes.

We observe strong correlations between visitation frequencies in the noise-free and noisy landscapes across all three transcription factors (new Supplementary Figure S35), indicating that evolutionary accessibility patterns are robust to realistic levels of experimental uncertainty. These analyses are described in the revised Results (lines 622–636) and in a new Supplementary Methods section (“Incorporation of experimental uncertainty into adaptive walks”).

Reviewer #2 (Public review):

The authors aim to investigate the ability of evolution to create strong transcription factor binding sites (TFBSs) de novo in E. coli. They focus on three global transcriptional regulators: CRP, Fis, and IHF, using a massively parallel reporter assay to evaluate the regulatory effects of over 30,000 TFBS variants. By analyzing the resulting genotype-phenotype landscapes, they explore the ruggedness, accessibility, and evolutionary dynamics of regulatory landscapes, providing insights into the evolutionary feasibility of strong gene regulation. Their experiments show that de novo adaptive evolution of new gene regulation is feasible. It is also subject to a blend of chance, historical contingency, and evolutionary biases that favor some peaks and evolutionary paths.

(1) Strengths of the methods and results:

The authors successfully employed a well-designed sort-seq assay combined with high-throughput sequencing to map regulatory landscapes. The experimental design ensures reliable measurement of regulation strengths. Their system accounts for gene expression noise and normalizes measurements using appropriate controls.

Comprehensive Landscape Mapping:

The study examines ~30,000 TFBS variants per transcription factor, providing statistically robust and thorough maps of the regulatory landscapes for CRP, Fis, and IHF. The landscapes are rigorously analyzed for ruggedness (e.g., number of peaks) and epistasis, revealing parallels with theoretical uncorrelated random landscapes.

Evolutionary Dynamics Simulations:

Through simulations of adaptive walks under varying population dynamics, the authors demonstrate that high peaks in regulatory landscapes are accessible despite ruggedness. They identify key evolutionary phenomena, such as contingency (multiple paths to peaks) and biases toward specific evolutionary outcomes.

Biological Relevance and Novelty:

The author's work is novel in focusing on global regulators, which differ from previously studied local regulators (e.g., TetR). They provide compelling evidence that rugged landscapes are navigable, facilitating de novo evolution of regulatory interactions. The comparison of landscapes for CRP, Fis, and IHF underscores shared topographical features, suggesting general principles of global transcriptional regulation in bacteria.

(2) Weaknesses of the methods and results:

Undersampling of Genotype Space:

While the quality filtering of the data ensures robustness, ~40% of the TFBS space remains uncharacterized. The authors acknowledge this limitation but could improve the analysis by employing subsampling or predictive modeling.

We thank the reviewer for raising this point. We agree that undersampling of genotype space is an important limitation of our dataset and that, in principle, subsampling or predictive modeling approaches could be used to address missing genotypes. We have now clarified in the manuscript why these approaches are not straightforward in the context of our analyses and why we did not pursue them here.

Although approximately 40% of TFBS genotypes were removed during the filtering step due to lack of reliable measurements, this filtering step was necessary to ensure robust estimation of regulation strength from sort-seq data. Importantly, random subsampling of the genotypes in our data set would not alleviate this limitation, because many of our key analyses—such as peak identification, quantification of epistasis, and assessment of evolutionary accessibility—require combinatorially complete local neighborhoods in genotype space. Subsampling would remove mutational neighbors from many neighborhoods, and thus further limit our ability to characterize landscape topology.

Predictive modeling approaches could, in principle, be used to infer missing genotypes and reconstruct more complete landscapes. However, developing, experimentally validating, and benchmarking such models would not only substantially expand the scope of an already long paper, it would also require additional assumptions about genotype–phenotype relationships that entail their own limitations. Our primary goal in this work was to provide the first large-

scale empirical *in vivo* regulatory landscapes for global bacterial transcription factors, comprising tens of thousands of experimentally measured variants. We view these empirical landscapes as a necessary foundation upon which predictive modeling and landscape completion can be built in future, complementary studies.

We have now revised the Discussion (lines 760-770) to explicitly articulate these points and to clarify that, while undersampling remains a limitation, it does not invalidate the landscape-level conclusions we draw from the combinatorially complete neighborhoods present in our data. There we also outline predictive modeling as an important directions for future work.

For a more detailed answer regarding subsampling and peak classification, please also see our response to comment (2) of Reviewer #1.

Simplified Regulatory Architecture:

The study considers a minimal system of a single TFBS upstream of a reporter gene. While this may have been necessary for clarity, this simplification may not reflect the combinatorial complexity of transcriptional regulation in vivo.

Point well taken. We have added paragraph to state explicitly that the system we use to study gene regulation is much simpler than most *in vivo* regulatory circuits (Discussion, lines 797-802)

Lack of Experimental Validation of Simulations:

The adaptive walks are based on simulated dynamics rather than experimental evolution. Incorporating in vivo experimental evolution studies would strengthen the conclusions. Although this is a large request for the paper, that would not prevent publication.

We thank the reviewer for this important point. We fully agree that *in vivo* experimental evolution would provide a valuable and complementary way to validate the evolutionary dynamics inferred from our simulations. However, we ask for the reviewer's understanding that adding experimental evolution to an (already long) paper would go far beyond the scope of our study.

Also, the goal of our study was not to reproduce evolutionary trajectories experimentally, but to characterize the structure of large empirical regulatory landscapes, and to use these landscapes as a data-driven basis for exploring evolutionary accessibility under well-defined population-genetic assumptions. The adaptive walks we employ are parameterized directly from experimentally measured genotype–phenotype maps, and incorporate established fixation probabilities. Such walks have been widely used to study evolutionary dynamics on empirical landscapes when experimental evolution is not tractable, because it would involve tens of thousands of genotypes that represent small mutational targets and would thus take a long time to evolve.

An additional issue related to the feasibility of experimental evolution is that performing *in vivo* experimental evolution for the regulatory landscapes analyzed here would require tracking large populations across a combinatorially vast TFBS space, while simultaneously measuring regulatory phenotypes for thousands of evolving lineages, which is currently not experimentally feasible. This is another reason why simulation-based approaches have been the standard method for linking large-scale empirical landscapes to evolutionary dynamics in both theoretical and experimental studies.

Furthermore, our conclusions are intentionally framed at the level of statistical and landscape-wide properties (e.g., accessibility of high peaks, contingency, and evolutionary bias), rather than at the level of specific mutational trajectories. As such, they do not rely on

the precise reproduction of any single evolutionary path, but on aggregate patterns that are robust to reasonable variation in population-genetic parameters.

In sum, we do not view experimental evolution as essential for the conclusions we draw, but as an important and exciting direction for future work that may be enabled by the landscapes we have experimentally mapped.

Impact on the Field:

This study advances our understanding of adaptive landscapes in gene regulation and offers a critical step toward deciphering how global regulators evolve de novo binding sites. The findings provide foundational insights for synthetic biology, evolutionary genetics, and systems biology by highlighting the evolutionary accessibility of strong regulation in bacteria.

Utility of Methods and Data

The sort-seq approach, combined with landscape analysis, provides a robust framework that can be extended to other transcription factors and systems. If made publicly available, the study's data and code would be valuable for researchers modeling transcriptional regulation or studying evolutionary dynamics.

Additional Context:

The study builds on a growing body of work exploring regulatory evolution. For instance, recent studies on local regulators like TetR and AraC have revealed high ruggedness and epistasis in TFBS landscapes. This study distinguishes itself by focusing on global regulators, which are more biologically complex and influential in bacterial gene networks. The observed evolutionary contingency aligns with findings in other biological systems, such as protein evolution and RNA folding landscapes, underscoring the generality of these evolutionary principles.

Conclusion:

The authors successfully mapped the genotype-phenotype landscapes for three global regulators and simulated evolutionary dynamics to assess the feasibility of strong TFBS evolution. They convincingly demonstrate that ruggedness and epistasis, while prominent, do not preclude the evolution of strong regulation. Their results support the notion that gene regulation evolves through a blend of chance, contingency, and evolutionary biases.

This paper makes a significant contribution to the understanding of regulatory evolution in bacteria. While minor limitations exist, the authors' methods are robust, and their findings are well-supported. The work will likely be of broad interest to researchers in molecular evolution, synthetic biology, and gene regulation.

We thank the reviewer for their thorough evaluation and for their supportive opinion of this paper.

Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

(1) Line 28 (Abstract): "Landscape ruggedness does not prevent the evolution of strong regulation, because more than 10% of evolving populations can attain one of the highest peaks." I did not find this interpretation very convincing; only 10% of populations being able to achieve strong regulation sounds to me like ruggedness DOES impede adaptation in the vast majority of cases.

We thank the reviewer for this thoughtful comment and agree that our original phrasing in the Abstract overstated this conclusion. We did not intend to imply that landscape ruggedness has only a minor effect on adaptation. On the contrary, our results clearly show that ruggedness strongly constrains evolutionary outcomes and prevents the majority of evolving populations from reaching the globally highest regulatory peaks. We have therefore toned down the wording in both the Abstract and the Discussion (lines 670-679) to reflect this more accurately. For example, in the abstract we now state

“Nonetheless, evolutionary simulations show that ~10% of evolving populations can reach a peak of strong regulation, a proportion that is significantly greater than in comparable random landscapes.”

In the discussion we state:

“... Specifically, our evolutionary simulations show that 10% of populations with a size typical of *E. coli* reach one of the highest peaks. This percentage is significantly higher than in randomized landscapes (Supplementary Methods 9; Supplementary Figure S30)”

Our intended interpretation was more limited: namely, that ruggedness does not fully preclude the evolution of strong regulation. In highly rugged landscapes with extensive sign epistasis—whose topological properties approach those of uncorrelated random landscapes—the a priori expectation is that access to the strongest peaks could be vanishingly rare or effectively impossible under Darwinian evolution. In this context, observing that a non-negligible fraction of populations (on the order of 10%) can reach one of the highest peaks suggests that strong regulation remains evolutionarily attainable, even though it is far from guaranteed.

Motivated by the reviewer’s suggestion, we also added a null-model analysis that makes this point more explicitly and quantitatively. Specifically, we constructed randomized landscapes by permuting regulation-strength values across genotypes while preserving the experimentally sampled genotype network topology and all parameters of the evolutionary simulations (Supplementary Methods 9, “Randomized landscape null model for peak accessibility”). We then repeated the adaptive-walk simulations on these shuffled landscapes. This null model provides an expectation for peak accessibility in landscapes with identical sampling, neighborhood structure, and evolutionary dynamics, but without genotype–phenotype correlations.

Using this null model, we find that the fraction of populations that reach high peaks in the empirical landscapes is substantially higher than expected by chance alone (new Supplementary Figure S30; Results, lines 504–516). Specifically, across the three transcription factors, empirical landscapes exhibit on average a ~3-fold higher accessibility of high regulatory peaks than shuffled landscapes. This comparison does not weaken the conclusion that ruggedness strongly impedes adaptation; rather, it shows that the structure of the measured genotype–phenotype landscapes enables greater accessibility of strong regulation than would be expected in equally rugged but unstructured landscapes.

In response to the reviewer’s concern, we have revised the abstract and main text to avoid the phrase “does not prevent” and to more accurately convey this balance between constraint and accessibility. We now emphasize that ruggedness strongly constrains adaptation, while still allowing access to strong regulatory peaks at rates that exceed null expectations. (Discussion, lines 512-516). For example, in the discussion we state:

“... In sum, rugged regulatory landscapes strongly constrain evolutionary trajectories, yet do not render the evolution of strong regulation vanishingly rare. Instead, strong regulatory phenotypes remain evolutionarily attainable at levels that exceed null expectations, even though they are reached by only a minority of evolving populations.”

We believe that the revised wording, together with the added null-model analysis more faithfully represents our results and strengthens the quantitative interpretation of accessibility in these landscapes.

(2) Line 123: I found the explanation of the plasmid system and the accompanying SI figures (Figures S1 and S2) confusing in terms of how many plasmids there were. In particular, the Figure S1 graphics show the plasmid specifically with CRP but the text in the graphic and in the caption refers to the plasmid pCAW-Sort-Seq-V2 (which, according to Table S1, isn't that just the base plasmid without any TF?). Figure S2 also shows the plasmid with CRP and does specify pCAW-Sort-Seq-V2-CRP-CRP0 in the graphic, but then the caption refers again only to the base plasmid pCAW-Sort-Seq-V2. I recommend the authors clarify these items for readers who might want to reproduce or build upon their system. In particular, I recommend the main text explain more explicitly that they generate three versions of this plasmid (one for each TF), and then on the backgrounds of each of those three plasmids, a whole library with all the binding site variants.

We thank the reviewer for pointing out this lack of clarity. We agree that the original description of the plasmid system and the accompanying Supplementary Figures S1 and S2 could be confusing with respect to how many plasmids were used and how they differ.

To clarify the experimental design, we start from a common backbone plasmid, pCAW-Sort-Seq-V2, which contains all shared regulatory and reporter elements but does not encode any transcription factor. From this backbone, we generated three distinct TF-specific plasmids, each carrying one of the transcription factors studied here—CRP, Fis, or IHF—resulting in pCAW-Sort-Seq-V2-CRP, pCAW-Sort-Seq-V2-Fis, and pCAW-Sort-Seq-V2-IHF. On the background of each TF-specific plasmid, we then constructed a complete library of plasmids containing all variants of the corresponding TF binding site cloned upstream of the reporter gene.

We have revised the main text to explicitly describe this plasmid hierarchy and library construction strategy and to clarify that three TF-specific plasmids were generated prior to TFBS library construction (Results, Landscape mapping section; lines 159–193). In addition, we have redesigned Supplementary Figures S1 and S2 to facilitate understanding of the plasmid system. Specifically, these figures now clearly distinguish between the base plasmid backbone and the TF-specific plasmid derivatives. Also, the plasmid names shown in the graphics and captions are now consistent with those listed in Supplementary Table S1. Upon final publication, we will also deposit the sequences of all plasmids in Addgene to further facilitate reproducibility.

(3) Line 135: Can the authors clarify whether these TFs are essential in these media conditions and, if not, why? I was expecting them to be so given the core functions of these TFs as described in the Introduction, but then Figure S3 appears to show that all knockouts are viable.

We thank the reviewer for raising this important point and apologize for the lack of clarity in the original version of the manuscript. The transcription factors CRP, Fis, and IHF are not essential for viability under the growth conditions used in this study, but they are important for optimal growth and cellular fitness, consistent with their roles as global regulators.

Under our experimental conditions, single-gene knockout strains (Δ crp, Δ fis, and Δ ihf) are viable but exhibit slower growth dynamics compared to the wild-type strain, reflecting impaired regulation of core cellular processes (Supplementary Figure S3). This behavior is consistent with previous work showing that many global transcriptional regulators in *E. coli* are conditionally essential or strongly fitness-affecting, rather than absolutely essential under standard laboratory conditions.

Importantly, while single knockouts remain viable, double mutants involving these global regulators are not viable, indicating substantial functional redundancy and network-level essentiality among global transcription factors. This explains why each TF can be studied individually in isolation, while combinations of deletions cannot be maintained.

We have now clarified this point in the Results section by explicitly stating that the knockout strains show reduced growth rates but reach comparable cell densities during late exponential or early stationary phase, the growth phase at which all measurements were performed (Results, Landscape mapping section; lines 185–193). This clarification reconciles the apparent discrepancy between the biological importance of these transcription factors discussed in the Introduction and the viability of the single-knockout strains shown in Supplementary Figure S3.

(4) Lines 141 and 227: The authors appear to refer to two different citations for different versions of RegulonDB (refs. 47 and 66). Did they actually use both versions for different purposes (if so, why?), or is this a typo?

We thank the reviewer for noticing this inconsistency. We did not use two different versions of RegulonDB. The two separate references were an error. We have now corrected this by using a single, consistent RegulonDB citation in both locations.

(5) Line 166 (Figure 1 caption): I think 2^8 here should be 4^8 .

Thank you. We have corrected “ 2^8 ” to “ 4^8 ” in the Figure 1 caption.

(6) Figure 2 Are the distributions in Figure 2a (regulation strengths across all TFBSs in the libraries) equivalent to the distributions in Figures S4–S6 (direct fluorescence readout from cell sorting), just transformed from fluorescence to regulation strength? If so I think that would be helpful to clarify, perhaps in the captions to Figures S4–S6 so that it's clear these contain the same information.

No. Figures S4–S6 and Figure 2a do not show the same distributions. Figures S4–S6 display the raw fluorescence distributions obtained from cell sorting, whereas Figure 2a shows regulation strengths (S), which are derived quantities computed from these fluorescence data. Specifically, regulation strength is calculated as a weighted average over fluorescence bins using the sequencing read distribution for each TFBS (see Methods, “Regulation strengths”).

To clarify this relationship, we have revised the main text (lines 201–203 and Figure 1b–c), to explicitly state how regulation strengths (S) were calculated.

(7) Figure 2b: Can the authors label each logo/frequency matrix with its corresponding TF name in the graphic itself? I think this is only implied in the caption.

We have updated Figure 2b to label each sequence logo / frequency matrix directly in the graphic with its corresponding transcription factor name (CRP, Fis, or IHF), in addition to mentioning these names in the caption. This change clarifies the figure and makes the TF identity immediately apparent to the reader.

(8) Lines 290 and 298 (Figure 2 caption): The labels for panels b and c appear to be swapped in the caption.

We thank the reviewer for pointing this out. The labels for panels b and c in the Figure 2 caption were indeed swapped. This has now been corrected.

(9) Line 379: There is a missing period at the end of this line.

We have added the missing period at the end of this line.

| (10) Line 400 (Figure 3 caption): *There is a missing subtitle for panel c in the caption for this figure (all other panels seem to have bolded subtitles in their captions).*

We have added the missing subtitle for panel c in the Figure 3 caption to match the formatting of the other panels.

| (11) Line 583: *There is a missing period after "Methods 7.5)".*

We have added the missing period after “Methods 7.5)”.

| (12) Line 641: *"All three landscapes highly rugged" should probably be "All three landscapes are highly rugged".*

We have corrected the sentence to read “All three landscapes are highly rugged.”

<https://doi.org/10.7554/eLife.103774.2.sa0>