Title: Entangled adaptive landscapes facilitate the evolution of gene regulation by exaptation

Author list: Cauã Antunes Westmann[1,2], Leander Goldbach[1,2], Andreas Wagner[1,2,3*]

Affiliations:
[1] – Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, Zurich CH-8057, Switzerland
[2] – Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, 1015 Lausanne, Switzerland
[3] – The Santa Fe Institute, Santa Fe, NM 87501, USA
* - Corresponding author: andreas.wagner@ieu.uzh.ch

Abstract

Exaptation, the co-option of existing traits for new functions, is a central process in Darwinian evolution. However, the molecular changes leading to exaptations remain unclear. Here, we investigated the potential of bacterial transcription factor binding sites (TFBSs) to evolve exaptively, examining three global transcription factors (TFs) from *Escherichia coli*: CRP, Fis, and IHF. Using a massively parallel reporter assay, we mapped three combinatorially complete adaptive landscapes, encompassing all intermediate sequences between three pairs of strong TFBSs for each TF. Our results revealed that these landscapes are smooth and navigable, with a monotonic relationship between mutations and their impact on gene regulation. Starting from a strong TFBS for one of our TFs, Darwinian evolution can create a strong TFBS for another TF through a small number of individually adaptive mutations. Notably, most intermediate genotypes are prone to transcriptional crosstalk – gene regulation mediated by both TFs. Because our landscapes are smooth, Darwinian evolution can also easily create TFBSs that show such crosstalk whenever it is adaptive. Our study presents the first in vivo evidence that new TFBSs can evolve exaptively through multiple small and adaptive mutational steps. It also highlights the importance of regulatory crosstalk for the diversification of bacterial gene regulation.

Keywords: adaptive landscapes, navigability, gene regulation, bacteria, massively parallel reporter assay

Introduction:

Exaptation – the repurposing of an existing trait for a new function – is an important route towards new traits in evolution. This process spans various biological levels, from the organismal—like bird feathers, initially thermoregulatory structures that became indispensable for flight [1,2] — to the molecular level [3,4], such as the evolution of eye crystallins from metabolic enzymes [5,6].

However, both the molecular changes and the adaptive value of intermediates in exaptation remain largely unknown. Transcriptional regulation provides a relevant system for studying exaptation due to its central role in life and adaptation[7–10]. In this process, a transcription factor protein (TF) binds a DNA sequence called a transcription factor binding site (TFBS), regulating gene expression[11–13]. Through exaptation, a TFBS initially recognized by one TF can evolve to be recognized by another TF. For example, during the evolution of vertebrate αA-crystallin, a TFBS known as a heat shock element was co-opted into a TFBS for the TF Pax6 to drive crystallin expression in the eye lens [5]. Another example is the co-option of regulatory regions of genes from transposable elements in the development of the mammalian uterus[14 15,16].

To understand the origin of such exaptations in molecular detail requires an understanding of the adaptive landscapes on which TFBSs evolve[17,18]. An adaptive landscape is an analogue to a physical landscape in which individual locations correspond to a genotype, such as the DNA sequence of a TFBS. The elevation at each location corresponds to a quantitative phenotype, such as the ability of a TFBS to regulate gene expression, which depends on how strongly the TF binds the TFBS.

The evolutionary process by which a strong ancestral TFBS for some TF (TF1) becomes exapted as a strong TFBS for another TF (TF2) usually requires a series of nucleotide changes (mutational paths). We call an adaptive landscape whose evolutionary paths correspond to transitions between TFBSs (or other traits) an *exaptation landscape*. A fundamental question about such a landscape is whether there exist evolutionary paths towards the derived (exapted) TFBS that are *accessible* to natural selection – paths in which each mutational step increases regulation strength for the derived TFBS.

Any evolutionary path to a new TFBS may also harbor intermediate TFBSs that can regulate gene expression through both TF1 and TF2, a phenomenon called transcriptional crosstalk [19–22]. Theoretical estimates suggest that 1-10% of all prokaryotic genes and 10-50% of eukaryotic genes may be subject to crosstalk [21] and its adaptive role is still debated [19–22]. Unfortunately, no pertinent empirical data is available.

Here, we experimentally map the exaptation landscape of three pairs of prokaryotic TFs in *Escherichia coli* [11,23,24], namely CRP (*cAMP receptor protein*[25]), Fis (*Factor for Inversion Stimulation*[26]) and IHF (*Integration Host Factor*) [27]. These TFs come from distinct families [28], have different biological roles, and are known as global regulators because they regulate hundreds of genes [11,23 23,29]. Such binding promiscuity increases the likelihood of transcriptional crosstalk.

We ask for all three pairs of these TFs (CRP-Fis, CRP-IHF, Fis-IHF) whether it is possible for a strong TFBS recognized by one TF to evolve into a strong TFBS for another TF through accessible mutational paths. We called such strong TFBSs "wild-type" sequences (WT_{TF}). Using a combinatorially complete library[30], comprising all possible mutants between wild-type

pairs, we measured the strength of gene regulation conferred by each mutant through a green fluorescent protein (GFP) reporter system.

We show that all three TF exaptation landscapes are smooth. Each has two global peaks of strong TFBSs, one for each TF, which permits exaptation through adaptive evolution. We also show that all intermediate TFBSs between the wild-type sequences for each pair of TFs are subject to crosstalk. That is, they can mediate regulation by both TFs to some extent. This extensive crosstalk can hinder or facilitate the exaptation of prokaryotic TFBSs, depending on its adaptive value.

Results

A sort-seq approach for mapping exaptation landscapes

To measure how strongly the TFBSs we study regulate gene expression, we redesigned a plasmid-based system from a previous study [31] (**Figure 1a, Supplementary Figures S1-S2, Supplementary Methods 2-3**). This system allowed us to express our TFs in a controlled manner, and to quantify the ability of a TFBS to be recognized by the TF through its effects on the regulation of the GFP reporter gene (**Figures 1b-c**). For brevity, we refer to this ability as the strength of a TFBS or its *regulation strength*.

To study the evolution of new TFBSs on three exaptation landscapes, we performed the following procedure for each of our three pairs of *E. coli* TFs. We started from strong TFBSs for each TF (named WT_{CRP}, WT_{Fis}, WT_{IHF}) in each TF pair, and aligned these reference TFBSs to determine the smallest number of nucleotide changes needed to convert one into the other (**Figure 1d, top panel, Supplementary Figure S3, Supplementary Tables S2-S3**). We then synthesized a library of DNA sequences containing each sequence on each mutational path between these strong binding sequences (**Figure 1d, bottom panel**).

Subsequently, we measured the regulation strengths of each sequence in this library to both TFs. In other words, each sequence becomes associated with two quantities, reflecting its ability to mediate gene expression regulation by each of the two TFs. To this end, we employed a massively parallel reporter assay called "sort-seq"[31] (**Figure 1e**, **see Methods**, **Supplementary Methods 7.2**). The landscapes we studied for TF pairs CRP-Fis, CRP-IHF,

and Fis-IHF comprised 256, 128, and 128 distinct TFBSs, respectively (**Figure 1f, Supplementary Figure S3 Supplementary Table S3**).

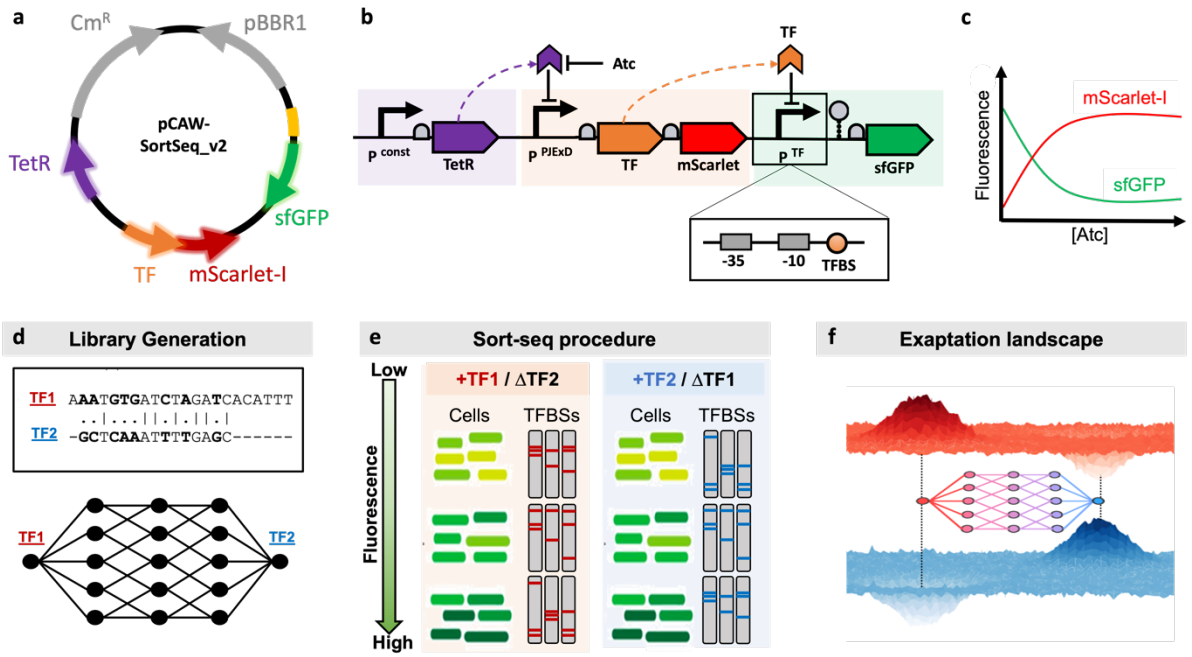

Figure 1. Experimental workflow. a-b. The reporter system and its regulatory modules. We redesigned a previously developed plasmid[31] (**Supplementary Methods 2-3**) that encodes a chloramphenicol resistance gene as a selective marker (CmR, grey, right-oriented arrow) and features a broad-host low-copy replication origin (pBBR1, grey, left-oriented arrow). This plasmid harbors three modules interconnected through regulatory interactions. The first module comprises the constitutively expressed *tetR* gene (purple). The TetR regulator controls the second regulatory module (red box) through the *plTetO1* promoter[32]. This second module encompasses a bicistronic operon, comprising both the TF of interest (red) and a reporter *mScarlet-I* gene[33] (red). In the third module (green box), the TF interacts with its cognate TFBS or a variant from the appropriate library. This interaction results in the transcriptional repression of the fluorescent reporter gene *sfgfp*[34] through steric hindrance, blocking RNA polymerase activity. The addition of anhydrotetracycline (Atc) alleviates repression by TetR, prompting the expression of the TF of interest, which leads to repression of *sfgfp*. To safeguard against the unintended effects of 5'- untranslated regions (UTRs) created by some TFBS library variants on the expression of the *sfgfp* mRNA, we used the RiboJ[35,36] transcriptional insulator (hairpin structure, in grey). Ribosome binding sites (RBSs) are highlighted as grey semi-circles. **c. Basic reporter dynamics.** As the TetR inducer (Atc) levels increase, more TetR regulator is sequestered from its binding site, allowing the expression of the TF and the mScarlet-I reporter (red). Increasing TF concentration causes a decrease in green fluorescence, because the TF causes increasing repression of the sfGFP reporter. **d. Library generation.** Alignment of two strong-affinity ("wild-type") TFBSs for individual TFs, indicating positions (dots) at which the two aligned sites differ from each other. We created a binding site library by varying the nucleotides at all nucleotide positions that differ between the two sites. The library comprises 2^n sequences, where n is the number of positions at which the sites differ (See **Supplementary Methods 4, Supplementary Table S3** and **Supplementary Figure S3** for library details). **e. Transformation of the reporter system harboring the library into two distinct bacterial strains.** One strain expresses TF1 (+TF1) but not TF2 (ΔTF2), and the other TF2 (+TF2) but

not TF1(ΔTF1). We then performed fluorescence-activated cell sorting (FACS) independently for each strain, separating cells (green rectangles) into specific bins according to their fluorescence. High-throughput sequencing of the library variants from each bin reveals the mutational differences in each variant relative to the reference TFBS, indicated as red and blue bars inside grey cylinders (TFBS), corresponding to data from the two TF-producing strains. The strength of regulation (binding) conveyed by each binding site is measured from the site's read count across bins (**Supplementary Methods 7.2**). **f. Mapping an Exaptation landscape.** Each TFBS is characterized by two regulation strengths, one for each TF. The regulation strength of each genotype defines an adaptive landscape for each TF (shown as the central-colored network, with the color gradient representing the regulation strengths for TF1, red, and TF2, blue). These can be interpreted as two entangled adaptive landscapes (red and blue above and below the network), where a single sequence is mapped to two phenotypes.

TFBS libraries for pairs of global regulators vary widely in their regulation strengths

Because each of the TFs we studied is also encoded in the *E. coli* genome, their simultaneous expression from a plasmid and the genome would not allow us to discriminate their individual regulation strengths for sequences in our library. We thus performed our experiments in *E. coli* strains in which the chromosomal copy of one TF was deleted (Δ*crp*, Δ*fis*, Δ*ihf*, see **Supplementary Table S4**). By doing so, we were able to isolate the effect of one TF from that of the other TF in a pair. We observed no substantial growth differences between the deletion strains (**Supplementary Figure S4**).

To validate our reporter system, we studied the regulatory strength of our three wild-type TFBSs (WT_{CRP}, WT_{Fis}, WT_{IHF}, **see Supplementary Table S2**) in the presence of their cognate TFs (**Supplementary Figure S5**). We found that each wild-type TFBS can indeed strongly regulate expression through its cognate TF (**Supplementary Figure S5**). More importantly, we found negligible crosstalk between the wild-type TFBSs (**Supplementary Figure S5**).

After these preliminary experiments, we characterized the interactions between each TF pair and their respective libraries (**Figure 2**). As an example, to characterize the CRP-Fis library, we studied this library in two reporter plasmids and two *E. coli* strains, one plasmid encoding

183 CRP (pCRP) in a Δ*fis* strain in which Fis is not expressed chromosomally (pCRP-Δ*fis*, **Figure**
184 **2a**), the other encoding Fis (pFis) in a Δ*crp* strain in which CRP is not expressed
185 chromosomally (pFis-Δ*crp*, **Figure 2a**). We applied the same strategy to each library (**Figures**
186 **2a-c,** see also **Methods section "Sort-seq procedure for dissecting the effects of individual**
187 **TFs"**).
188
189 Sort-seq (**see Methods, Supplementary Methods 5, Supplementary Figures S6-S8**) allowed
190 us to characterize how strongly each TFBS variant in each of our libraries can regulate gene
191 expression in response to each of the two appropriate TFs (**Figures 2d-f**). Briefly, we
192 characterized the strength of regulation conveyed by each TFBS variant by averaging the
193 sequencing read counts of each sequence across fluorescence bins and normalizing them by
194 their respective wild-type regulatory strength (**Supplementary Methods 7.2**). The Pearson's
195 correlation coefficient R for pairwise associations of sequencing read counts for replicates of
196 each library ranged from 0.923 to 0.992 (**Supplementary Figures S9-S11**), indicating a high
197 technical reproducibility. The distribution of these scores is approximately normal
198 (**Supplementary Figure S12**), with many sequences regulating expression with intermediate
199 strength by both TFs, and few sequences regulating expression strongly by only one TF.
200
201 We further validated these observations by selecting ten sequences from each landscape
202 (**Supplementary Tables S6-S8**) that cover a wide range of measured regulatory strengths, and
203 individually measured their effects on GFP expression in the presence of either TF1 or TF2.
204 The correlation between these direct measurements and sort-seq data is strong and nearly linear
205 for all pairs of TFs, with Pearson's correlation coefficients R ranging from -0.71 to -0.81
206 (**Methods, Supplementary Figure S13**).
207

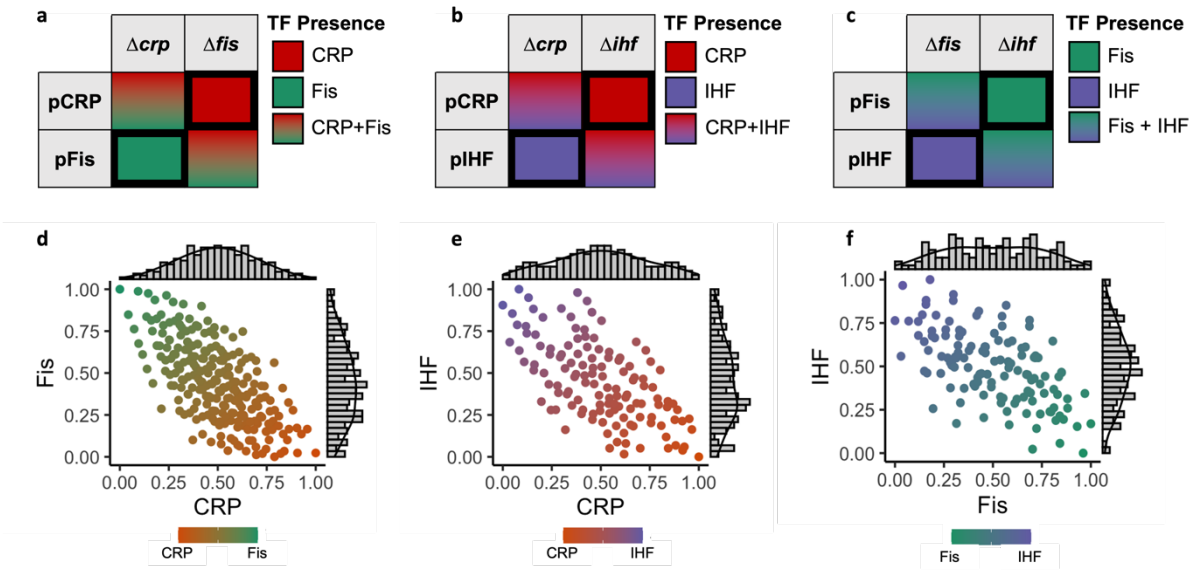

Figure 2. Characterization of regulation strengths for individual TFs. a-c. Specific plasmid-strain combinations ensure the presence of only one of two TFs during our measurements. Each table schematically represents the intracellular presence of TFs for combinations of a reporter plasmid encoding one of the TFs (rows) transformed into a mutant *E. coli* strain lacking the gene encoding one of the TFs (columns). Pure colors represent the presence of a single TF in the cell (CRP: red; Fis: green and IHF: purple), and color gradients represent the presence of both TFs. We avoided plasmid-strain combinations in which both TFs were present in the cell, as such combinations did not allow us to unambiguously measure the regulation conveyed by one TF's binding to each of the TFBSs in an exaptation landscape. The plasmid-strain combinations we used are highlighted with thick black borders. **d-f. Distribution of regulation strengths for three TFBS libraries.** Each panel shows data from a different TF1-TF2 library (d. CRP-Fis; e. CRP-IHF; f. Fis-IHF). Each circle represents a single TFBS variant in a landscape. The horizontal and vertical axes represent the regulation strengths of each variant in the respective landscape to the TF labelled on the axes. Pure colors represent the regulation strengths of each TF (CRP: Red; Fis: Green: IHF, blue), and intermediate colors represent binding by two TFs. The histograms on the top and right of each plot show the distribution of regulation strengths for each TF.

Exaptation landscapes are smooth and reveal abundant crosstalk

Next, we used our sort-seq-based data to map all three exaptation landscapes. We study each such landscape as a network of genotypes (TFBSs) where two genotypes are connected by an edge (straight line) if they differ by a single nucleotide (**Figure 3**). This data already contains important clues to the question of whether exaptation could occur by natural selection alone. The first clue is that the distribution of regulatory strength ratios changes smoothly with increasing distance from each wild-type TFBS (**Figures 3a-c**).

We quantified crosstalk for each TFBS as a normalized ratio of the regulatory strengths by TF1 and TF2 (**Figures 3d-f**, see Methods). While the wild-type TFBS is regulated predominantly by the cognate TF in any pair of TFs (**Figure 3d-f, Supplementary Figure S5**), all intermediate TFBSs show at least some ability to be regulated by the other TF in the pair (**Figures 3d-f**). This means that even single mutations away from the wild type can be favored by natural selection. A dimensionality reduction analysis (**Supplementary Figure S14, Supplementary Methods 7.6**), as well as an analysis of nucleotide frequency changes during TFBS evolution (**Supplementary Figure S15**), shows that some nucleotide changes are more important than others in driving changes of regulation strength during exaptation.

A second clue comes from an analysis of the landscapes' peaks – TFBSs for which all neighbouring genotypes regulate transcription less strongly than the peak itself. Any one such peak (different from the wild-type) may hinder exaptation, because an evolving population could become trapped on local suboptimal peaks with low fitness[17,37]. However, for each TF1-TF2 landscape, there are only two peaks, one for each TF, corresponding to the two wild-type sequences (**Figures 3a-c**).

Third, we also studied the fraction of evolutionary paths to a wild-type sequence that would be accessible to selection, i.e., a series of mutational steps in which each step increases regulation strength. In all three exaptation landscapes, most paths starting from the wild-type TFBS of TF1 and leading to the wild-type sequence of the TF2 are accessible (ranging from 65% to 100%, **Supplementary Figure S16)**. The reason why some paths to the peak are not accessible in two landscapes is that these landscapes show (reciprocal) sign epistasis (**Supplementary Methods 7.5**, **Supplementary Tables S6-S8** and **Supplementary Figure S16**). These are

forms of non-additive interactions among sequence variants, in which one or two individual mutations reduce regulation strength but both mutations together increase it. Specifically, in the CRP-Fis landscape, 4% of double mutants show (reciprocal) sign epistasis. In the CRP-IHF landscape, 11% of double mutants show (reciprocal) sign epistasis.

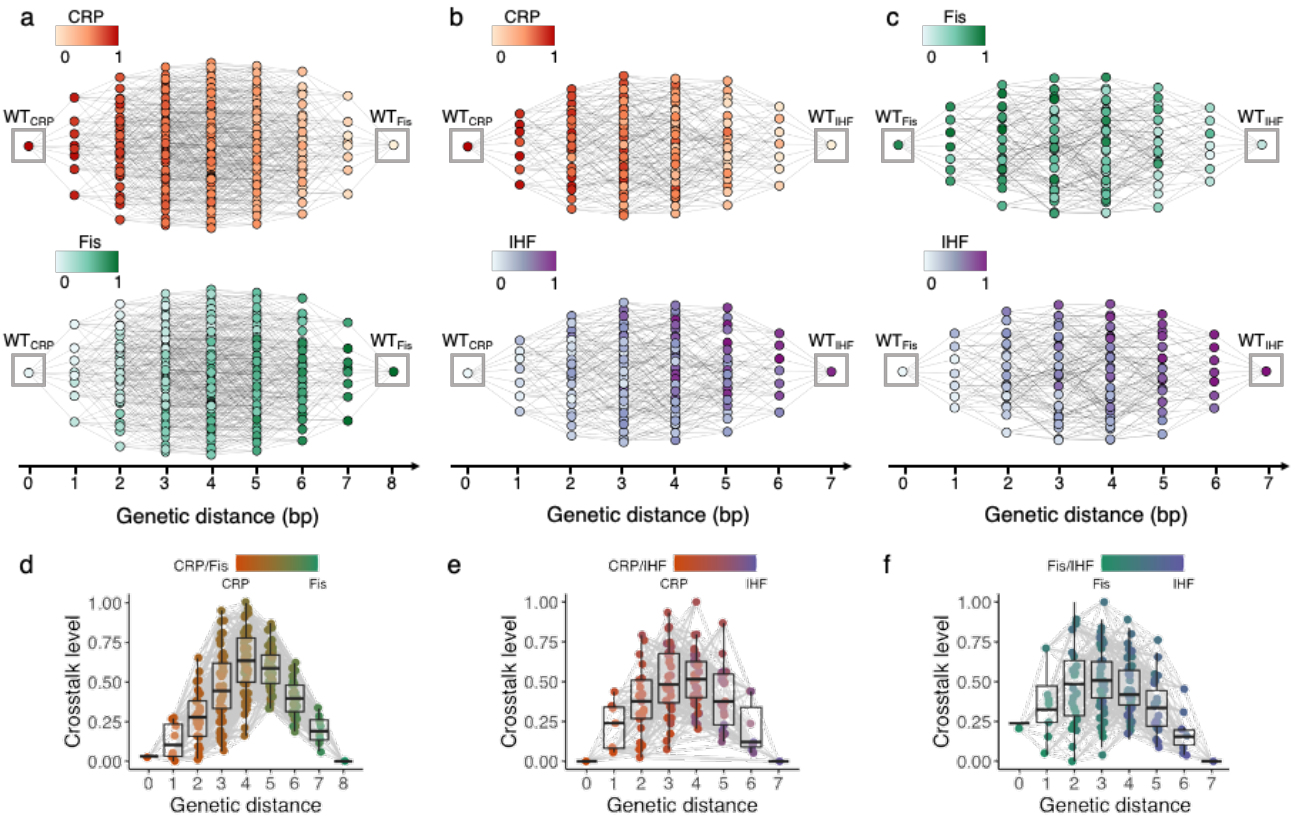

Figure 3. Exaptation landscapes are smooth with many intermediates subject to transcriptional crosstalk. Panels a-c depict exaptation landscapes as networks whose nodes (circles) represent binding site variants connected through single nucleotide changing mutations (straight lines). The x-axis of each panel shows the genetic distance (in number of nucleotide differences) from each TFBS variant to one of the wild-type sequences (at x = 0). Grey squares indicate WT sequences for each TF, which are peaks in their respective landscapes. **a. The CRP-Fis exaptation landscape.** Both panels represent the CRP-Fis exaptation landscape (N = 256) with nodes ordered by their distance to the CRP wild-type sequence. Colors represent regulation strengths of each TFBS variant to CRP (top panel, red heatmap) and to Fis (bottom panel, green heatmap). B**. The CRP-IHF exaptation landscape.** Both panels represent the CRP-IHF exaptation landscape (N = 128) with nodes ordered by their distance to the CRP wild-type sequence. Colors represent regulation strengths of each TFBS variant to CRP (top panel, red heatmap) and to IHF (bottom panel, purple heatmap). **c. The Fis-IHF exaptation landscape.** Both panels represent the Fis-IHF exaptation landscape (N = 128) with nodes ordered by their distance to the Fis wild-type sequence. Colors represent Fis (top panel, green heatmap) and IHF (bottom panel, purple heatmap) regulation strengths for each sequence. Light and dark colors in panels a-c represent

weak and strong binding, respectively. Note that all regulation strengths in panels a-c are normalized by the regulation strength of the wild-type sequence. **d-f. Crosstalk levels for binding sites in each exaptation landscape.** Each panel shows, for all TFBSs in an exaptation landscape, the ratio of binding to TF1 and TF2 (y-axis) as a function of the TFBSs genetic distance from the wild-type sequence (x-axis), which is a strong TFBS. Each circle symbolizes one genotype. Grey circles connect genotypes that differ in a single nucleotide. Boxplots indicate the distribution of crosstalk levels for each mutational distance. Each box covers the range between the first and third quartiles (interquartile range, IQR). The horizontal line within the box represents the median value, and whiskers span 1.5 times the IQR. Heatmap colors reflect the regulation strength ratio for each TFBS. That is, sequences binding predominantly to CRP, Fis or IHF are displayed in red, green, and purple, respectively. Intermediate colors indicate binding at intermediate strength to two TFs. Note that crosstalk levels are normalized by the maximum crosstalk value measured, resulting in a scale that ranges from 0 (no crosstalk) to 1 (maximum crosstalk). We generated networks with the *Gephi*[38] software using a combination of both the *3DForceAtlas* and *Network3DSplitter* algorithms.

In silico analysis suggests crosstalk between two closely related bacterial species

To study the exaptive evolution of TFBSs, one would ideally like to infer their evolutionary history with phylogenetic methods, an approach that has proven successful for eukaryotic TFBSs[39–42]. In order to explore if we could find genomic indications of TFBS exaptation in bacteria, we performed a simple analysis with the genomes of two closely related species, *Escherichia coli* K12 / MG1655 and *Salmonella typhimurium* LT2 (see **Supplementary Methods 7.7**). By comparing noncoding DNA sequences upstream of orthologous genes and operons between the two species, we found evidence of both *de novo* evolution and exaptive evolution in TFBSs for eight *E. coli* global regulators (CRP, Fis, IHF, ArcA, FNR, Cra, Lrp, Fur)[23]. We considered evidence for exaptive evolution when a TFBS for one TF within an intergenic region in one species overlapped with a TFBS for a different TF in the other species. Specifically, we observed an average gain of 3.2±0.3 (mean±s.d.) TFBSs per intergenic region through either *de novo* or exaptive evolution, with 34% of these gains potentially driven by exaptation. This observation is noteworthy given that the species shared a most recent common ancestor approximately 100 million years ago[43] (**Supplementary Figure S17, Supplementary Methods 7.7**).

316 **Rapid exaptation when Darwinian evolution favors strong binding by one TF**

317 Motivated by the potential for naturally occurring TFBS exaptations in bacteria, we examined
318 evolutionary dynamics on our experimentally mapped landscapes, simulating population
319 evolution in the strong selection weak mutation (SSWM) regime (**Supplementary Methods**
320 **8**) [44–47]. We focused on this regime due to its relevance in large bacterial populations such as in
321 *E. coli* [48] (exceeding 10^8 individuals), where genetic drift is minimal, and even subtle mutations
322 are visible to natural selection [44–47]. Given the low probability of mutations at individual TFBS
323 nucleotide positions (2.2×10^{-10} per generation)[48], beneficial mutations typically reach fixation
324 before the next mutation arises, resulting in a monomorphic population –- concentrated at a
325 singular genotype on the landscape – that adapts through stepwise mutations, or an adaptive
326 walk [44–47 47,49.]

327

328 We simulated 10^4 adaptive walks per TF1-TF2 landscape, starting from each of the 10 strongest
329 TFBSs for TF1 (for a total of 10^5 walks). We defined each genotype's (TFBS) fitness by TF2
330 regulation strength, and fixation probabilities were based on population genetic theory [50,51],
331 which allows the calculation of fixation probabilities for fitness-affecting mutations as a
332 function of population size [50,51](**Supplementary Methods 8**). Additionally, we simulated
333 reverse-direction walks using the highest-regulation genotypes for TF1 as starting points,
334 incorporating known *E. coli* mutation biases [52,53] (**Supplementary Methods 8**) and adapting
335 evolution simulations for populations of 10^2 and 10^8 individuals.

336

337 In large populations (10^8), where selection is strong and drift is weak [54,55], all adaptive walks
338 reached the wild-type TFBS for TF2 or TF1 (CRP-Fis: **Figures 4a-b**; CRP-IHF:
339 **Supplementary Figures S18a-b**; Fis-IHF: **Supplementary Figures S19a-b**). Asymmetrical
340 path lengths, like CRP to Fis versus Fis to CRP (5.3±0.9 vs. 5±0.9 steps, mean±s.d., $p<0.001$,

Welch Two Sample t-test), likely result from sign epistasis [17,56,57] (**Supplementary Tables S9-S11**). Such an asymmetry also exists in the CRP-IHF landscape (**Supplementary Figures S18 c-f**). Despite such modest asymmetries, paths to each peak were nearly as short as the minimal genetic distance, reflecting landscape smoothness and navigability (CRP-Fis: **Figures 4e-f**; CRP-IHF: **Supplementary Figures S18e-f**; Fis-IHF: **Supplementary Figures S19e-f**). Contingency, defined as the dependence of a historical process on chance events[58–60], was evident, as multiple paths led to identical peaks (**Supplementary Figures S20-S21**). Lastly, in smaller populations ($N=10^2$), where selection is weaker and drift is stronger [54,55], populations showed reduced fitness (~1% reduction, $p<2.2\times10^{-16}$, One-Sample t-test, $n=10^5$, **Supplementary Figures S23a-f**) and delayed peak attainment by 10% relative to large populations ($p<2.2\times10^{-16}$, Welch Two Sample t-test, $n=2\times10^5$, **Supplementary Figures S23g-l**).

Additionally, we investigated the impact of deviations from the SSWM regime on our observations. When mutation rates are high or populations are even larger than we consider, populations are polymorphic most of the time [49,61,62]. In consequence, they experience clonal interference where the fittest among several simultaneously segregating mutations is likely to fixate [49,61,62]. High mutation rates and clonal interference, simulated via "greedy" adaptive walks [37,45,63,64] (**Supplementary Methods 8**), revealed no significant exaptation rate differences compared to the SSWM regime (**Supplementary Figures S24-S26**). In sum, these smooth, single-peaked landscapes enable rapid exaptation via selection, regardless of genetic drift or clonal interference.

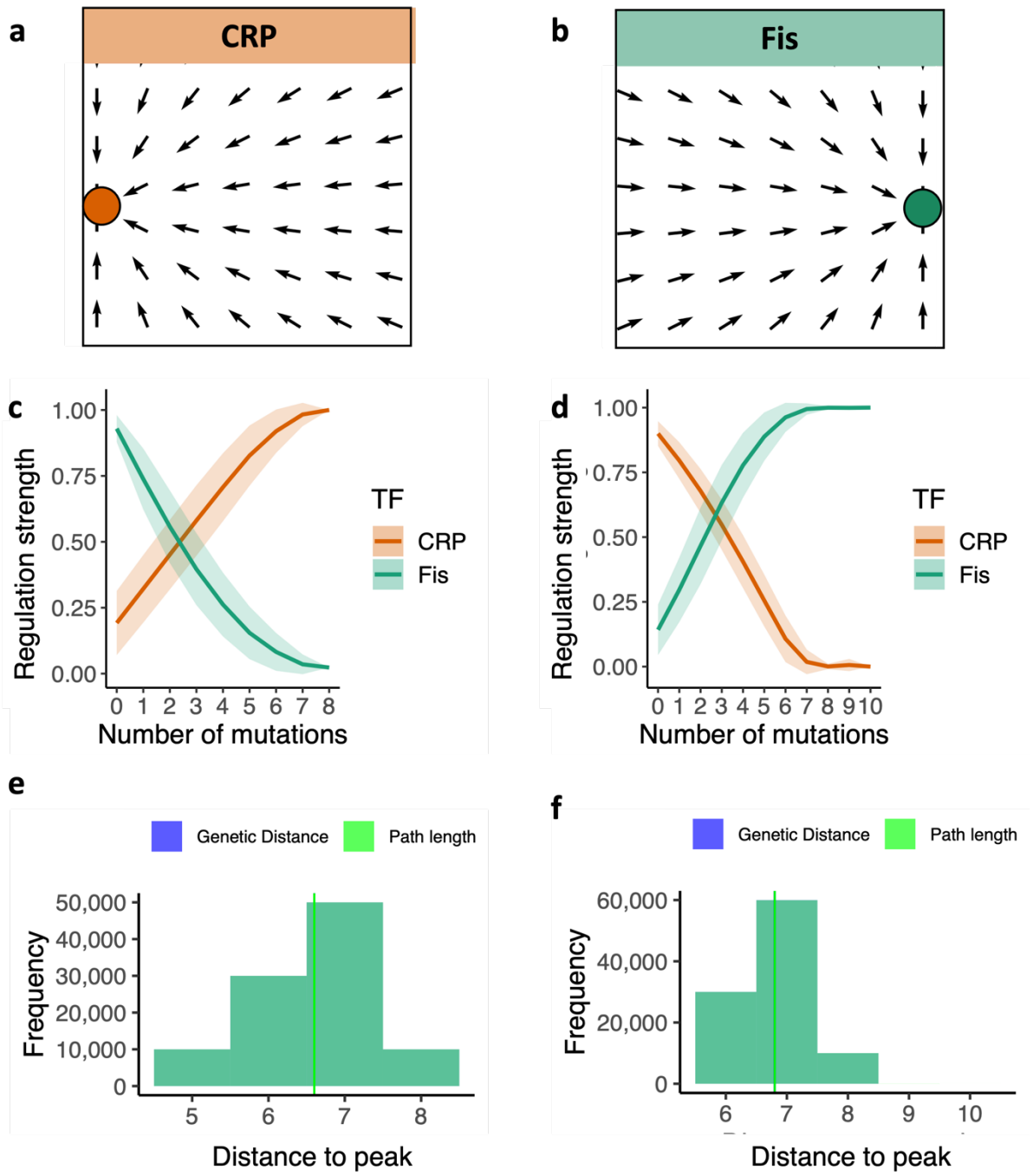

Figure 4. Rapid exaptation on the CRP-Fis landscape. All data presented is based on a population of 10^8 individuals. Data for each plot is based on 10^5 simulated adaptive walks, i.e., 10^4 adaptive walks for each of the 10 starting genotypes with the highest regulation strength for the opposite TF. **a-b. Vector field illustrations of adaptive walks in genotype space.** The illustrations show the direction of adaptive walks in genotype space when strong binding to a specific TF (CRP or Fis) is favored. Each panel represents a single TF landscape, with the x and y axes representing coordinates in genotype space. Each vector indicates the direction of an adaptive walk starting from a distinct genotype and ending at the colored circle peak ("attractor"), which signifies strong regulation by the TF. **a. Scenario 1: Adaptive trajectories favoring strong binding to CRP.** In the CRP landscape, adaptive walks favor strong binding to CRP, indicated by the orange circle (CRP peak). **b. Scenario 2: Adaptive trajectories favoring strong binding to Fis.** In the Fis landscape, adaptive walks favor strong binding to

Fis, indicated by the green circle (Fis Peak). **c-d. Adaptive walks readily attain peaks.** Each plot represents how regulation strengths (vertical axis) for both TFs (CRP in orange and Fis in green) change along the transversed path length (horizontal axis) from the initial sequence to the attained peak during an adaptive walk. Curves represent averages over all 10^5 adaptive walks, and shading indicates one standard deviation of regulation strengths. **c. Adaptive walks attain the CRP peak.** The data is based on 10^4 walks starting from each of the 10 strongest binding sites for Fis, favoring increasing CRP binding. **d. Adaptive walks attain the Fis peak.** Like panel (c), but for adaptive walks starting from the 10 strongest binding sites for CRP, favoring increasing Fis binding. **e-f. Evolutionary paths to a peak are not much longer than the minimal genetic distances.** Each composite histogram shows the distribution of path lengths (green) for all pairs of starting (non-peak) and ending (peak) genotypes during 10^5 adaptive walks. Blue histograms represent the distribution of genetic distances for each pair of non-peak and peak genotypes. The green and blue horizontal lines represent average path lengths and mutational distance, which are indistinguishable. We used a t-test of the null hypothesis that the means of these distance distributions are statistically indistinguishable. **e. Distribution of accessible path lengths and genetic distances to the CRP peak.** We observed no significant difference between mean genetic distance and path length (mean ± s.d.: 6.6 ± 0.7 for both genetic distance and path length; Welch Two Sample t-test: t = 31.973, df = 196,700, p-value = 0.7, $N_1 = 10^5$ genetic distances, $N_2 = 10^5$ path lengths). **f. Distribution of accessible path lengths and genetic distances to the Fis peak.** We observed no significant difference between mean genetic distance and path length (mean ± s.d.: 6.9 ± 0.8; Welch Two Sample t-test: t = 35.113, df = 196,700, p-value = 0.82, $N_1 = 10^5$ genetic distances, $N_2 = 10^5$ path lengths).

In silico analysis suggests extensive crosstalk between TFBSs in the *E. coli* genome

We next focused on crosstalk between our three TFs, and first asked whether crosstalk might exist between them in the *E. coli* genome. To this end, we used Position-Weight Matrices (PWMs). Each such matrix encodes information on experimentally verified TFBSs for any one TF and represents the probability that any one of the four DNA nucleotides occurs at each specific position of these TFBSs[65,66]. PWMs are well-established computational tools to score how strongly an unknown DNA sequence may be bound by a given TF [65,66]. For our three TFs and the TFBSs in our libraries, these computational PWM scores are correlated with our experimentally measured regulation strengths (mean Pearson R = 0.56±0.1, mean±s.d., N= 3, see **Supplementary Methods 7.8 and Supplementary Figure S27**). We thus used PWMs together with 756 experimentally validated *E. coli* TFBSs from the RegulonDB database[28] for our three TFs, to predict the potential for crosstalk between these TFs in the *E. coli* genome. This potential is indeed very high (**Figure 5a, Supplementary Methods 7.8**).

412 In our analysis of 756 TFBSs with predicted regulation strengths for both CRP and Fis using
413 their respective PWMs, 94% (706/756) of the sequences exhibited predicted regulation
414 strengths above 0.25 for each transcription factor (**Figure 5b)**. This threshold was established
415 as the minimal level for the presence of crosstalk. In the same dataset, 78% (591/756) and 82%
416 (619/756) of the TFBSs surpassed this threshold for the CRP-IHF (**Figure 5c**) and Fis-IHF
417 (**Figure 5d**) pairwise comparisons, respectively. Furthermore, 44% (332/756) of the sequences
418 demonstrated predicted regulation strengths above 0.5 for CRP and Fis, which is half of the
419 maximum value computable by PWMs (the PWM maximum value is normalized to 1). These
420 sequences were categorized as having the potential for intermediate to high crosstalk.
421 Additionally, within the same dataset, 8% and 9% of the TFBSs exceeded the 0.5 threshold for
422 CRP-IHF and Fis-IHF pairwise comparisons, respectively.
423
424 The observed negative associations across pairwise comparisons can be explained by the
425 incompatibility between nucleotide positions of the distinct PWMs. For instance, sequences
426 closer to the CRP maximum predicted value tend to have lower values for IHF, and vice versa.
427 A similar pattern was observed for Fis and IHF. This could also suggest a selection pressure
428 against high crosstalk levels in genomic sequences, favoring sequences with low to medium
429 crosstalk levels, as observed in our analysis.
430
431 A computational comparison of PWMs for all TFs with a PWM available for *E. coli* also
432 suggests that this high potential for crosstalk is the rule rather than the exception for *E. coli*
433 TFs (**Supplementary Figure S28, Supplementary Methods 7.9**). Moreover, the comparison
434 between PWMs of global regulators shows that Fis and IHF PWMs are closer together than
435 CRP-Fis or CRP-IHF (**Supplementary Figure S29**).
436

Rapid exaptation also occurs when selection favors crosstalk between two TFs

Motivated by these observations of the potential for crosstalk in vivo and in silico, we speculated that crosstalk may at least sometimes be adaptive. We thus asked how readily crosstalk could evolve adaptively from crosstalk-free TFBSs in our exaptation landscapes.

Under the SSWM regime, we performed 10^4 random walks per genotype across each landscape, totaling N = 256 x 10^4, N = 128 x 10^4, and N = 128 x 10^4 for the CRP-Fis, CRP-IHF, and Fis-IHF landscapes, respectively. Fitness was defined as the sum of regulation strength for both TFs (CRP-Fis: **Figures 5e-g**; CRP-IHF: **Supplementary Figures S30 a,c,e**; Fis-IHF: **Supplementary Figures S30 b,d,f**; **Supplementary Methods 8**). Both CRP-Fis and Fis-IHF landscapes remained single-peaked, with intermediate regulatory strengths for each TF (CRP-Fis: 0.63 for CRP, 0.75 for Fis; Fis-IHF: 0.68 for Fis, 0.76 for IHF). However, CRP-IHF had two peaks: 87% of walks reached the first peak (CRP: 0.68, IHF: 0.68), and 13% reached the second (CRP: 0.87, IHF: 0.32).

Consistent with our previous simulations without crosstalk, paths to each peak were as short as the minimal genetic distance, highlighting each landscape's smoothness and navigability. For instance, in CRP-Fis, evolutionary paths and mutational distances were not statistically different (Welch Two Sample t-test, p-value = 0.85, **Figures 5 f-g**). Similar results were observed for CRP-IHF (**Supplementary Figures S30 a,c,e**) and Fis-IHF (**Supplementary Figures S30 b,d,f**).

In small populations (N=10^2), genetic drift had variable effects across landscapes. In CRP-Fis, drift increased CRP regulation by 10% and decreased Fis by 14% (**Supplementary Figure S31a**). For CRP-IHF, the most accessed peak reduced CRP by 13% and IHF by 8%. In Fis-

IHF, drift increased Fis by 8% and reduced IHF by 13% (**Supplementary Figure S31 d**). Drift did not impact peak attainment time, except for the most accessed peak in the CRP-IHF landscape, where the average number of mutations increased to reach such peak increased from 3.4 to 5 (45% increase, $p<2.2 \times 10^{-16}$, Welch Two Sample t-test, $n = 256 \times 10^4$, **Supplementary Figures S31 e-h**). We also simulated deviations from the SSWM regime using adaptive "greedy" walks, confirming that clonal interference did not alter exaptation rates, as high-fitness peaks were reached with similar steps as in the SSWM regime (**Supplementary Figure S32**). In sum, selection-driven evolution for crosstalk progresses rapidly on these single-peaked landscapes, with minimal interference from genetic drift or clonal interference.

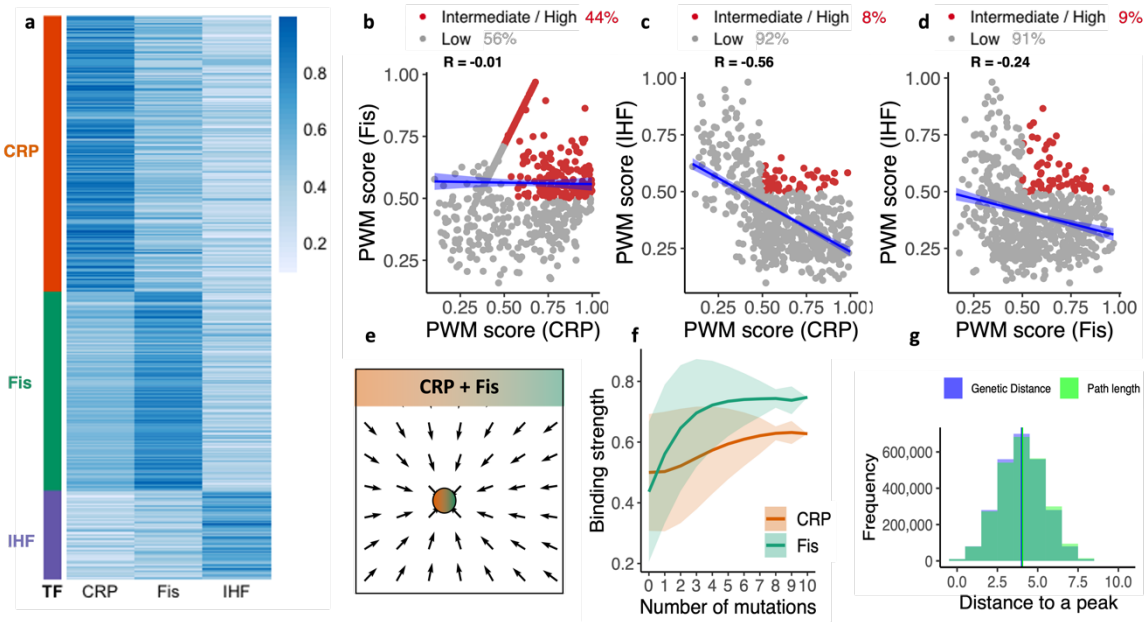

Figure 5. Transcriptional crosstalk and its consequences for evolutionary dynamics. a. PWM scores for experimentally characterized TFBSs from the *E. coli K-12* genome suggest ample crosstalk between TFs CRP, Fis, and IHF. The heatmap is based on 756 TFBS sequences in the *E. coli K-12* genome, as curated in the RegulonDB[67] database. Each row corresponds to one TFBS. Specifically, the dataset includes 370 TFBSs for CRP (orange in left-most column), 267 for Fis (green), and 119 for IHF (purple). Each of the remaining columns quantifies the predicted binding strength (PWM score) for each of the 756 TFBSs with respect to the transcription factors CRP, Fis, and IHF, as indicated below each column. We normalized predicted binding strength by the maximum possible score for each TF, resulting in a scale from 0 to 1, as indicated by the color gradient in the heatmap color legend. We clustered the rows of the heatmap using a combination of Euclidean distance and hierarchical clustering with complete-linkage to reveal patterns of similarity in predicted binding strengths among the sequences. **b-d. Association between predicted binding strengths for CRP, Fis and IHF and their genomic *E. coli* TFBS.** We computed for each pair of TFs the statistical association

between predicted binding strength (CRP-Fis, CRP-IHF, Fis-IHF). In the scatter plots, red circles represent TFBSs whose average predicted binding strength for both TFs is higher than 0.5, indicating intermediate to high crosstalk. Grey circles represent TFBSs with an average predicted binding strength below this threshold, indicating low crosstalk. The percentage of sequences with intermediate to high crosstalk (red circles) and low crosstalk (grey circles) is shown at the top of the plots in red and gray. The blue line is a linear regression line (blue shading: 95% confidence interval; R: Pearson correlation coefficient.). **b. Association between CRP and IHF.** No significant association between CRP and Fis predicted binding strengths (R = -0.01 and t-statistic p = 0.71), which is due to the clear distinction between sparsely scattered points and the straight line involving some sequences with very similar predicted binding strength values. **c. Association between CRP and IHF.** R = -0.56 and t-statistic p-value $< 2 \times 10^{-6}$. **d. Association between Fis and IHF.** R = -0.24 and t-statistic p-value $< 2 \times 10^{-6}$. **e-g. Rapid evolution of crosstalk in the CRP-Fis landscape. e. Schematic of evolution towards TFBSs with high crosstalk.** The vector field illustrates the direction of adaptive walks in genotype space when strong binding towards a CRP and Fis is favored. Each vector indicates the direction of an adaptive walk starting from a distinct genotype and ending at a peak ("attractor", colored circle) strongly binding both TFs. **f. Adaptive walks readily attain TFBSs binding to both CRP and Fis with intermediate strengths.** We performed 10^4 adaptive walks starting from each of the 256 TFBS variants in the CRP-Fis landscape, with selection favoring strong binding for both CRP and IHF. The panels show the regulation strength (vertical axis) for CRP (orange) and Fis (green) plotted against the transversed path length of an evolving sequence to the attained peak (horizontal axis). Curves represent averages over all 256 x 10^4 adaptive walks and shading indicates one standard deviation of regulation strengths. **g. The distribution of accessible path lengths and genetic distances to peaks.** The green and blue horizontal lines show the average length of traversed paths during 256 x 10^4 walks between pairs of non-peaks and peaks, and the respective mutational distances between such pairs, respectively. The means of the distributions are statistically indistinguishable (3.31±1, mean ± s.d., for both distance metrics, Welch Two Sample t-test t = -5.5782, df = 2029535, p-value = 0.7, N_1= 256 x 10^4, N_2= 256 x 10^4).

518 **Discussion:**

519 We experimentally demonstrated that exaptation landscapes for three pairs of *E. coli* TFs (CRP-
520 Fis, CRP-IHF, Fis-IHF) are smooth, with only two peaks, one for each regulator. Despite the
521 landscape smoothness, sign epistasis in certain landscapes restricts evolutionary paths.
522 Simulations show that strong TFBSs for TF1 can evolve into strong TFBSs for TF2 with few
523 mutational steps, highlighting smooth, accessible transitions. Neither strong genetic drift nor
524 clonal interference significantly alters this conclusion.

525

526 Each of these global regulators controls hundreds of *E. coli* genes and often co-regulate
527 complex *E. coli* promoters[68–71]. In specific environments, it may be advantageous for some of
528 these genes to become rewired and regulated by a different global regulator. While gene
529 regulatory rewiring is common across bacteria [72–76], documented cases of TFBS exaptation are
530 rare, likely due to challenges in observing them across species amid frequent horizontal gene
531 transfers and recombination events [72,77–79]. Our results suggest regulatory exaptation could be
532 a frequent mechanism in bacterial gene regulation diversification.

533

534 Although theoretical models of TFBS exaptation predominate [17,80–82], recent experimental work
535 has shed light on this process [17,80–82]. For example, a synthetic biology study engineering
536 overlapping TFBSs to accommodate up to three TFs [83,84]. Another study investigated the
537 potential of two distinct TFs (Lambda CI and P22 C2) to bind a small library of TFBS
538 mutants[85]. The study found that Lambda CI could recognize new TFBSs more easily when
539 starting from P22 C2's TFBS than the reverse [85]. A possible reason for this asymmetry may be
540 a greater incidence of epistasis in the landscape of one TF, like we observed in our landscapes
541 (**Figures 4e-f**, **Supplementary Figures S18-S19**).

542

In contrast to smooth adaptive landscapes, rugged (multi-peaked) landscapes can hinder Darwinian evolution by entrapping populations on local suboptimal peaks, and thus prevent a population's ascendance of the highest fitness peak[17]. With some exceptions[31,86,87], empirical landscapes of gene regulation are smooth[17]. The smoothness observed likely results from low epistasis among mutated nucleotide positions, as evidenced by alignments of TFBSs with less critical informational positions.

A growing body of theoretical studies suggests that transcriptional crosstalk is a global phenomenon in nature[19–22,88], and our empirical and in silico analysis further support this hypothesis. We observed extensive crosstalk potential for intermediate genotypes across our landscapes. However, despite such potential, the influence of crosstalk on exaptation is still poorly understood. On the one hand, Darwinian evolution requires that all intermediates of an evolutionary process are functional, and crosstalk can facilitate the evolution of intermediate genotypes if regulation by both TFs is beneficial. Crosstalk can be adaptive when a gene needs to be expressed in multiple environments and by different TFs [19,22,88,89]. Such adaptive crosstalk can simplify gene regulation and coordinate cellular responses across diverse environments. Furthermore, in the compact genomes of bacteria, where space is limited, transcriptional crosstalk can help reduce the size of regulatory regions [13,71,83,84,90,91], as evidenced by the overlap of many TFBSs in bacterial promoters [71,89]. On the other hand, crosstalk may cause conflict between adaptive functions and thus impair adaptive evolution when precise gene regulation is paramount [19–21]. Previous studies have shown that genes with numerous interaction partners (e.g., protein-protein or regulatory interactions) are less likely to undergo horizontal gene transfer, because the fortuitous regulation of many genes by an exogenous TF can reduce organismal fitness by globally rewiring the host's regulatory network[92–95].

These considerations motivated us to ask whether adaptive evolution on our landscapes could easily create crosstalk when such crosstalk is adaptive. The answer is yes, because our landscapes remained smooth under this selection regime, such that evolution rapidly finds TFBSs with intermediate strengths for two TFs and does so through short evolutionary paths of merely three mutations, on average. Our observation that most *E. coli* genomic TFBSs for CRP, Fis, and IHF exhibit the potential for low to intermediate crosstalk levels also suggests that although high crosstalk might be globally avoided, it can also be quickly tuned up or down, depending on its adaptive value.

Our study has some limitations. First, focusing on gene regulation and crosstalk,we did not measure fitness directly. However, other studies suggest a link between strong gene regulation and high fitness [81,96–99], which may not always hold universally. Second, we assumed positive selection for increasing regulatory strengths as shown for both microbes and humans[96,97,99,100], though weak TFBSs can play important regulatory roles [101,102]. Given our smooth landscapes, weak regulation could still evolve effectively under Darwinian evolution. Third, our system adopted a simple regulatory architecture, while native promoter architectures can involve multiple TFs and combinatorial regulation [70,71]. Such complex promoter architectures could potentially modulate crosstalk levels and should be further investigated. Lastly, we examined only a small subset of the TFBS sequence space. A broader sequence space may present more rugged, challenging landscapes for Darwinian evolution.

Within these limitations, we have shown that exaptive evolution in gene regulation can be easily accomplished, because the three exaptation landscapes we study are smooth. This perspective allows us to reconceptualize TFBS evolution as more complex than traditionally viewed, involving navigation through a space of interconnected landscapes where each

mutational step can influence the regulatory strength of multiple TFs. Examples of regulatory exaptation[6,15,103–105] and the existence of crosstalk in the wild[88,106] suggest that this may be a general principle of exaptation in gene regulation. To further corroborate this assertion, it will be necessary to study more transcription factors, measure fitness and crosstalk directly, explore a larger sequence space, and do so in the chromosomal context in which TFs operate.

599 **Methods**

600

601 The Supplementary methods contain extended details of experimental procedures and data
602 analysis.

603

604 **Media and reagents.** To prepare SOB medium, we mixed 25.5g from the solid medium stock
605 (VWR J906) with 960 ml of purified water and subjected the resulting suspension to
606 autoclaving. To prepare SOC medium, we dissolved 20 ml of 1 M D-glucose (Sigma G8270)
607 and 20 ml of 1 M magnesium sulfate (Sigma 230391) in 960 ml of the pre-prepared SOB
608 solution. For LB medium, we combined 25g of solid medium stock (Sigma-Aldrich L3522)
609 with 1 liter of purified water and then autoclaved it. To prepare M9 minimal medium, we
610 diluted M9 minimal salt sourced from Sigma (M6030) in distilled water as per the
611 manufacturer's guidelines, autoclaved the solution, and added 0.4% glucose (Sigma G8270),
612 0.2% casamino acid (Merk Millipore, 2240), 2 mM magnesium sulfate (Sigma 230391), and
613 0.1 mM calcium chloride (Sigma C7902). Where required, we supplemented growth media
614 with chloramphenicol (50 μg/mL), anhydrotetracycline (100 ng/mL, Cayman-chemicals
615 #10009542), and/or glucose (0.4% w/v final concentration). We prepared anhydrotetracycline
616 by diluting the dried chemical in absolute ethanol, from a stock concentration (1000X) of 100
617 μg/mL to a working concentration of 100 ng/mL.

618

619 **Strains and plasmids**

620 Bacterial strains and plasmds used in this work are listed in **Supplementary Tables S1- S2**.
621 We obtained electrocompetent *E. coli* cells (strain SIG10-MAX®) from Sigma Aldrich
622 (CMC0004). The genotype of this strain (**Supplementary Table S4**) is similar to DH5α
623 (Sigma Aldrich commercial information, see **Supplementary Table S4**). The strain is resistant

624 to the antibiotic streptomycin. Due to its high transformation efficiency, we used strain SIG10-
625 MAX® for molecular cloning and library generation.
626
627 We amplified plasmid libraries in DH5α, extracted their DNA, and transformed them into the
628 *E. coli* K-12 BW25113 wild-type strain and its derivative mutants to study gene regulation. We
629 obtained the *E. coli* wild type (BW25113) strain and its mutants, which harbor deletions of
630 CRP, IHF, or Fis global regulators from the KEIO collection [107], and used them for sort-seq
631 experiments. The design, genetic parts, and assembly of the plasmid vectors we used in this
632 study are available in the Supplementary material. All primers, TFBS sequences/libraries,
633 strains and plasmids are listed in **Supplementary Tables S1-S5**.
634
635 **Sort-seq procedure**.
636 We performed our sort-seq experiments as previously described[31]. We designed TF binding site
637 libraries in the Snapgene® software (snapgene.com). The libraries were synthesized by IDT
638 (Coralville, USA) as a single-stranded DNA Ultramer® of 140bp (4nmol). We resuspended
639 each library in nuclease-free distilled water and serially diluted it to a concentration of 50ng/uL.
640 We PCR-amplified library oligonucleotides, ligated them into the PCR-amplified pCAW-Sort-
641 Seq-V2 plasmid backbone (a derivative of pCAW-Sort-Seq [31], a low-copy number, multi-host
642 plasmid with the pBBR1 replication origin) by Gibson assembly, and transformed cells with
643 the resulting library after column-based purification with a commercial electroporation kit
644 (NEB #T1020L). We grew populations of *E. coli* cells hosting each library in LB medium
645 supplemented with 50 µg/mL chloramphenicol to saturation (overnight growth, 16 hours, 200
646 rpm, 37°C). After overnight growth, we diluted the overnight culture in LB medium
647 supplemented with 50 µg/mL chloramphenicol and anhydrotetracycline (100 ng/mL) in a 1:100
648 ratio (v/v) and grew the culture for 8 h until it had reached stationary phase. Before sorting, we

649 diluted 50μL of the culture in 1mL of cold Dulbecco's PBS (Sigma-Aldrich #D8537) in a 15
650 mL FACS tube. We performed FACS-sorting on a FACS Aria III flow cytometer (BD
651 Biosciences, San Jose, CA). We sorted and binned cells according to their GFP-mediated
652 fluorescence (FITC-H channel, 488nm laser, emission filter 502LP, 530/30). To this end, we
653 recorded the fluorescence of 10^6 cells and created 4 evenly spaced gates ("bins") on a binary
654 logarithmic ($\log_2$) scale spanning the observed range of fluorescence values (**see Methods,**
655 **Supplementary Figures S6-S8**). The number of cells sorted into each bin corresponded to the
656 fraction of the 10^6 cells recorded for that bin. Thus, the total number of sorted cells was 10^6.
657 We performed each sorting in triplicate from the same overnight culture. We isolated plasmids
658 from cells sorted into each bin (Qiagen, Germany) after overnight growth, and used a
659 polymerase chain reaction (PCR) to amplify the TFBS region from each plasmid for Illumina
660 sequencing. The distribution of library variants among the different bins allowed us to map
661 each individual genotype to a measure of how strongly it regulates gene expression.
662
663 The Supplementary material has additional details on library construction and sort-seq
664 procedures.
665
666 **Sort-Seq procedure for dissecting the effects of individual TFs**
667 To reduce the effects of extrinsic noise (cell-specific variation)[108] we followed common
668 practice[109–111] and normalized our *gfp* fluorescence values by *mscarlet* fluorescence values after
669 flow-cytometry assays (**see Supplementary Methods 5**). To dissect the distinct effects of
670 individual transcription factors within a shared pool of sequences (libraries), we employed a
671 sort-seq procedure that involved a combination of plasmids expressing specific TFs in specific
672 mutant bacterial strains. Below, we outline the key steps in this procedure for each TF pair.
673

674 **Sorting the CRP-Fis library**

675

676 **Characterizing the effect of Fis binding on gene regulation.** To study the influence of Fis
677 binding on our CRP-Fis library sequences independent from that of CRP binding, we proceeded
678 as follows: We cloned the CRP-Fis library onto the plasmid pCAW-Sort-Seq-V2-Fis, which
679 expresses the Fis transcription factor upon Atc induction. Next, we introduced this plasmid into
680 the *E.coli* Δ*crp* strain, in which the *crp* gene has been deleted (**Supplementary Table S4**,
681 **Figure 2a**). We then cultured the transformed bacterial cells overnight in LB medium
682 supplemented with Atc, which induces the expression of plasmid-encoded Fis within the
683 bacterial population. By doing so, we guaranteed complete absence of CRP and that Fis was
684 expressed. We then subjected the CRP-Fis library to the sort-seq procedure under these
685 conditions, allowing us to specifically examine the effects of Fis binding on our library
686 sequences.

687

688 **Characterizing the Effect of CRP Binding.** To study the impact of CRP binding on gene
689 regulation in isolation from that of Fis binding, we cloned the same CRP-Fis library on plasmid
690 pCAW-Sort-Seq-V2-CRP, expressing CRP in the Δ*fis* strain. We induced Fis expression with
691 Atc, ensuring that of the two transcription factors, only CRP is present. After overnight growth,
692 we sorted cells to measure the strength of CRP binding on the CRP-Fis library sequences.

693

694 **Sorting the CRP-IHF Library**. We adopted the same procedure we just described for the
695 CRP-Fis library, except that we cloned the CRP-IHF library into the plasmid pCAW-Sort-Seq-
696 V2-IHF, which expresses the IHF transcription factor upon Atc induction, and transformed this
697 library into a Δ*crp* strain. Briefly, to study the effect of IHF binding independent from that of
698 CRP binding, we performed sort-seq after library growth in LB supplemented with Atc. To

699 study the effect of CRP binding independent from that of IHF binding, we cloned the CRP-IHF
700 library into the plasmid pCAW-Sort-Seq-V2-CRP, which expresses the CRP transcription
701 factor upon Atc induction, and transformed this library into a Δ*fis* strain. We performed the
702 sort-seq procedure after library growth in LB supplemented only with Atc.
703
704 **Sorting the Fis-IHF Library**. To study the effect of Fis binding in isolation from that of IHF,
705 we cloned the Fis-IHF library into the plasmid pCAW-Sort-Seq-V2-Fis, and transformed the
706 resulting construct into the Δ*ihf* strain. Inducing the expression of Fis with Atc in this strain
707 ensured the expression of Fis in the absence of IHF. To study the effect of IHF binding in
708 isolation from that of Fis-binding, we cloned the same (Fis-IHF) library into the plasmid
709 pCAW-Sort-Seq-V2-IHF, and transformed the resulting plasmid library into the Δ*fis* strain. By
710 inducing the expression of IHF with Atc in this strain, we ensured the expression of IHF in the
711 absence of Fis.
712
713 **Replication of experiments**
714 Data from high-throughput methods to measure gene expression are intrinsically noisy. The
715 reasons include biological cell-to-cell expression variation [108], as well as technical [112–115]
716 variation (e.g. pipetting errors, equipment biases, limit of detection thresholds etc.). To account
717 for such variability, we performed sort-seq in three replicates, i.e., we performed three
718 independent sorting procedures for the same overnight culture.
719
720 **Regulation strengths**
721 Due to gene expression and measurement noise, individual TFBS variants in a sort-seq
722 experiment usually appear in more than a single bin, and their read count (frequency) varies
723 among bins[116–119]. Following established practice[31,118,120], we used a weighted average of these

724 frequencies for each variant to represent the mean expression level driven by the variant. To
725 facilitate interpretation, we converted this expression level into a regulation strength relative to
726 the wild type binding site (regulation strength of one) for that specific TF. Values below one
727 indicate weaker TF binding (higher GFP expression), whereas values above one indicate
728 stronger regulation strength (lower GFP expression).
729
730 **Validating regulation strengths with plate reader measurements**
731 To further validate our regulation strengths thus inferred, we chose 10 DNA binding sites that
732 cover a wide range of measured regulation strengths. We cloned these sequences into the
733 appropriate vector pCAW-Sort-Seq-V2-TF (TF: Fis or IHF) and transformed them into the
734 appropriate mutant strain to ensure that only Fis or IHF are expressed. We picked individual
735 colonies and grew them overnight (16 hours, 37°C, 220 rpm) in liquid LB supplemented with
736 50 µg/mL of chloramphenicol and anhydrotetracycline. We diluted the cultures to 1:10 (v/v) in
737 cold Dulbecco's PBS (Sigma-Aldrich #D8537) to a final volume of 1 mL. We transferred 200
738 µl of the diluted cultures into individual wells in 96-well plates and measured GFP fluorescence
739 (emission: 485nm/excitation: 510nm, bandpass: 20nm, gain: 50) as well as the optical density
740 at 600nm (OD_{600}) as an indicator of cell density. We then normalized fluorescence by the
741 measured OD_{600} value to account for differences in cell density among cultures and compared
742 the obtained ratios to the previously inferred regulation strengths for the 10 selected variants.
743 We performed all such measurements in biological and technical triplicates (three colonies per
744 sample and three wells per colony, respectively).
745
746

Data availability: The DNA sequencing data generated in this study have been deposited in the NCBI database (BioProject) under the accession code: PRJNA1162486. The flow cytometry and plate reader data generated in this study have been deposited in the Zenodo public repository and are accessible publicly via the following DOI: 10.5281/zenodo.14008581.

Code availability: The computer code generated in this study has been deposited in the Zenodo public repository and is publicly accessible via the following DOI: 10.5281/zenodo.14008581

References

1. Prum, R. O. & Brush, A. H. The Evolutionary Origin And Diversification Of Feathers. *Q Rev Biol* **77**, 261–295 (2002).
2. Pan, Y. *et al.* The molecular evolution of feathers with direct evidence from fossils. *Proceedings of the National Academy of Sciences* **116**, 3018–3023 (2019).
3. Wagner, A. Information Theory Can Help Quantify the Potential of New Phenotypes to Originate as Exaptations. *Front Ecol Evol* **8**, (2020).
4. Brosius, J. Exaptation at the molecular genetic level. *Sci China Life Sci* **62**, 437–452 (2019).
5. Cvekl, A. *et al.* Evolutionary Origins of Pax6 Control of Crystallin Genes. *Genome Biol Evol* **9**, 2075–2092 (2017).
6. Gavelis, G. S., Keeling, P. J. & Leander, B. S. How exaptations facilitated photosensory evolution: Seeing the light by accident. *BioEssays* **39**, 1600266 (2017).
7. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**, 206–216 (2007).
8. Nghe, P. *et al.* Predicting Evolution Using Regulatory Architecture. *Annu Rev Biophys* **49**, 181–197 (2020).
9. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59–69 (2012).
10. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170 (2016).
11. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol* **14**, 638–650 (2016).
12. Browning, D. F. D. F. & Busby, S. J. W. S. The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**, 57–65 (2004).
13. Struhl, K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**, 1–4 (1999).
14. Lynch, V. J. *et al.* Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell Rep* **10**, 551–561 (2015).
15. De Souza, F. S. J., Franchini, L. F. & Rubinstein, M. Exaptation of transposable elements into novel Cis-regulatory elements: Is the evidence always strong? *Molecular Biology and Evolution* vol. 30 1239–1251 Preprint at https://doi.org/10.1093/molbev/mst045 (2013).
16. Sundaram, V. & Wysocka, J. Transposable elements as a potent source of diverse *cis* -regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190347 (2020).
17. Aguilar-Rodríguez, J., Payne, J. L. & Wagner, A. A thousand empirical adaptive landscapes and their navigability. *Nat Ecol Evol* **1**, 0045 (2017).
18. Bendixsen, D. P., Collet, J., Østman, B. & Hayden, E. J. Genotype network intersections promote evolutionary innovation. *PLoS Biol* **17**, e3000300 (2019).
19. Grah, R. & Friedlander, T. The relation between crosstalk and gene regulation form revisited. *PLoS Comput Biol* **16**, e1007642 (2020).
20. Friedlander, T., Prizak, R., Barton, N. H. & Tkačik, G. Evolution of new regulatory functions on biophysically realistic fitness landscapes. *Nat Commun* **8**, (2017).
21. Friedlander, T., Prizak, R., Guet, C. C., Barton, N. H. & Tkačik, G. Intrinsic limits to gene regulation by global crosstalk. *Nat Commun* **7**, 12307 (2016).
22. Wagner, A. Adaptive gene misregulation. *Genetics* **217**, (2021).

23. Martínez-Antonio, A. *et al.* Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* **6**, 482–489 (2003).
24. Visweswariah, S. S. & Busby, S. J. W. 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 (2015).
25. Kolb, A., Busby, S., Buc, H., Garges, S. & Adhya, S. TRANSCRIPTIONAL REGULATION BY cAMP AND ITS RECEPTOR PROTEIN. *Annu Rev Biochem* **62**, 749–797 (1993).
26. Nash, H. A. & Robertson, C. A. Purification and properties of the Escherichia coli protein factor required for lambda integrative recombination. *J Biol Chem* **256**, 9246–53 (1981).
27. Koch, C. & Kahmann, R. Purification and properties of the Escherichia coli host factor required for inversion of the G segment in bacteriophage Mu. *Journal of Biological Chemistry* **261**, 15673–15678 (1986).
28. Salgado, H. *et al.* RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res* **52**, D255–D264 (2024).
29. Schmidt, A. *et al.* The quantitative and condition-dependent Escherichia coli proteome. *Nat Biotechnol* **34**, 104–110 (2016).
30. Reidys, C. M. & Stadler, P. F. Combinatorial Landscapes. *SIAM Review* **44**, 3–54 (2002).
31. Westmann, C. A., Goldbach, L. & Wagner, A. The highly rugged yet navigable regulatory landscape of the bacterial transcription factor TetR. *bioRxiv* 2023.08.25.554764 (2023) doi:10.1101/2023.08.25.554764.
32. Lutz, R. & Bujard, H. 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 (1997).
33. Bindels, D. S. *et al.* mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nature Methods 2016 14:1* **14**, 53–56 (2016).
34. Pédelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79–88 (2006).
35. Lou, C., Stanton, B., Chen, Y. J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol* (2012) doi:10.1038/nbt.2401.
36. Vlková, M., Morampalli, B. R. & Silander, O. K. Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to cis- and trans-changes in genetic background. *Microbiologyopen* **10**, e1232 (2021).
37. Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**, 11–45 (1987).
38. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third International AAAI Conference on Weblogs and Social Media* 361–362 (2009) doi:10.1136/qshc.2004.010033.
39. Koch, C. *et al.* Inference and Evolutionary Analysis of Genome-Scale Regulatory Networks in Large Phylogenies. *Cell Syst* **4**, 543-558.e8 (2017).
40. Thompson, D., Regev, A. & Roy, S. *Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution*. *Annual Review of Cell and Developmental Biology* vol. 31 (2015).
41. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of hormone-receptor complexity by molecular exploitation. *Science (1979)* **312**, 97–101 (2006).
42. Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).

43. Lawrence, J. G. & Ochman, H. Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A* **95**, 9413–9417 (1998).
44. Bank, C., Matuszewski, S., Hietpas, R. T. & Jensen, J. D. On the (un)predictability of a large intragenic fitness landscape. *Proc Natl Acad Sci U S A* **113**, 14085–14090 (2016).
45. Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56**, 1317–1330 (2002).
46. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature 2022 603:7901* **603**, 455–463 (2022).
47. Gillespie, J. H. Molecular Evolution Over the Mutational Landscape. *Evolution (N Y)* **38**, 1116–1129 (1984).
48. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics 2016 17:11* **17**, 704–714 (2016).
49. Li, J., Amado, A. & Bank, C. Rapid adaptation of recombining populations on tunable fitness landscapes. *Mol Ecol* (2023) doi:10.1111/MEC.16900.
50. Kimura, M. On the Probability of Fixation of Mutant Genes in a Population. *Genetics* **47**, 713–719 (1962).
51. Crow, J. and Kimura, M. *An Introduction to Population Genetics Theory*. (The Blackburn Press, 2009).
52. Lee, H., Popodi, E., Tang, H. & Foster, P. L. 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**, (2012).
53. Horton, J. S. & Taylor, T. B. Mutation bias and adaptation in bacteria. *Microbiology (N Y)* **169**, (2023).
54. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc of the 6th International Congress of Genetics* Preprint at (1932).
55. Weissman, D. B., Desai, M. M., Fisher, D. S. & Feldman, M. W. The rate at which asexual populations cross fitness valleys. *Theor Popul Biol* **75**, 286–300 (2009).
56. Payne, J. L. & Wagner, A. The Robustness and Evolvability of Transcription Factor Binding Sites. *Science (1979)* **343**, (2014).
57. Khalid, F. *et al.* Genonets server-a web server for the construction, analysis and visualization of genotype networks. *Nucleic Acids Res* **44**, W70–W76 (2016).
58. Louis, A. A. 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 (2016).
59. Blount, Z. D., Lenski, R. E. & Losos, J. B. Contingency and determinism in evolution: Replaying life's tape. *Science (1979)* **362**, (2018).
60. Vermeij, G. J. Historical contingency and the purported uniqueness of evolutionary innovations. *Proceedings of the National Academy of Sciences* **103**, 1804–1809 (2006).
61. Melissa, M. J., Good, B. H., Fisher, D. S. & Desai, M. M. Population genetics of polymorphism and divergence in rapidly evolving populations. *Genetics* **221**, (2022).
62. Stolyarova, A. V. *et al.* Complex fitness landscape shapes variation in a hyperpolymorphic species. *Elife* **11**, (2022).
63. Park, S.-C., Neidhart, J. & Krug, J. Greedy adaptive walks on a correlated fitness landscape. *J Theor Biol* **397**, 89–102 (2016).
64. ORR, H. A. A Minimum on the Mean Number of Steps Taken in Adaptive Walks. *J Theor Biol* **220**, 241–247 (2003).
65. Zhao, Y. & Stormo, G. D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* (2011) doi:10.1038/nbt.1893.
66. Stormo, G. D. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nat Rev Genet* **11**, 751–760 (2010).

67. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res* (2019) doi:10.1093/nar/gky1077.
68. Monteiro, L. M. O., Sanches-Medeiros, A., Westmann, C. A. & Silva-Rocha, R. Unraveling the Complex Interplay of Fis and IHF Through Synthetic Promoter Engineering. *Front Bioeng Biotechnol* **8**, 510 (2020).
69. Monteiro, L. M. O. *et al.* Emergent properties in complex synthetic bacterial promoters. *ACS Synth Biol* acssynbio.7b00344 (2017) doi:10.1021/acssynbio.7b00344.
70. Barnard, A., Wolfe, A. & Busby, S. Regulation at complex bacterial promoters: How bacteria use different promoter organizations to produce different regulatory outcomes. *Curr Opin Microbiol* **7**, 102–108 (2004).
71. Rydenfelt, M., Garcia, H. G., Cox, R. S. & Phillips, R. The influence of promoter architectures and regulatory motifs on gene expression in Escherichia coli. *PLoS One* **9**, 1–31 (2014).
72. Oren, Y. *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences* **111**, 16112–16117 (2014).
73. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–845 (2008).
74. Molina, N. Genome Evolution and Regulatory Network Structure in Bacteria. (2010).
75. Lozada-Chavez, I., Janga, S. C. & Collado-Vides, J. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res* **34**, 3434–3445 (2006).
76. Babu, M. M., Balaji, S. & Aravind, L. General Trends in the Evolution of Prokaryotic Transcriptional Regulatory Networks. in *Book* vol. 3 66–80 (KARGER, Basel, 2007).
77. Sakoparnig, T., Field, C. & van Nimwegen, E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *Elife* **10**, 1–61 (2021).
78. Maddamsetti, R. & Lenski, R. E. Analysis of bacterial genomes from an evolution experiment with horizontal gene transfer shows that recombination can sometimes overwhelm selection. *PLoS Genet* **14**, e1007199 (2018).
79. Arnold, B. J., Huang, I. T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology* 1–13 Preprint at https://doi.org/10.1038/s41579-021-00650-4 (2021).
80. Wagner, A. Information Theory Can Help Quantify the Potential of New Phenotypes to Originate as Exaptations. *Front Ecol Evol* **8**, 457 (2020).
81. Berg, J., Willmann, S. & Lässig, M. Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* **4**, 564–567 (2004).
82. Doniger, S. W. & Fay, J. C. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* **3**, 0932–0942 (2007).
83. Guazzaroni, M.-E. E. & Silva-Rocha, R. Expanding the Logic of Bacterial Promoters Using Engineered Overlapping Operators for Global Regulators. *ACS Synth Biol* **3**, 666–675 (2014).
84. Amores, G. R., Guazzaroni, M. E. & Silva-Rocha, R. Engineering Synthetic cis-Regulatory Elements for Simultaneous Recognition of Three Transcriptional Factors in Bacteria. *ACS Synth Biol* **4**, 1287–1294 (2015).
85. Igler, C., Lagator, M., Tkačik, G., Bollback, J. P. & Guet, C. C. Evolutionary potential of transcription factors for gene regulatory rewiring. *Nat Ecol Evol* **2**, 1633–1643 (2018).
86. Lagator, M., Igler, C., Moreno, A. B., Guet, C. C. & Bollback, J. P. Epistatic interactions in the arabinose cis-regulatory element. *Mol Biol Evol* **33**, 761–769 (2016).
87. Lagator, M., Paixão, T., Barton, N. H., Bollback, J. P. & Guet, C. C. On the mechanistic nature of epistasis in a canonical cis-regulatory element. *Elife* **6**, 1–16 (2017).

88. Taylor, T. B., Shepherd, M. J., Jackson, R. W. & Silby, M. W. Natural selection on crosstalk between gene regulatory networks facilitates bacterial adaptation to novel environments. *Curr Opin Microbiol* **67**, 102140 (2022).
89. Hermsen, R., Tans, S. & Ten Wolde, P. R. Transcriptional regulation by competing transcription factor modules. *PLoS Comput Biol* **2**, 1552–1560 (2006).
90. Westmann, C. A., Guazzaroni, M.-E. & Silva-Rocha, R. Engineering Complexity in Bacterial Regulatory Circuits for Biotechnological Applications. *mSystems* **3**, e00151-17 (2018).
91. van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation. *Microbiology and Molecular Biology Reviews* **73**, 481–509 (2009).
92. Wellner, A., Lurie, M. N. & Gophna, U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol* (2007) doi:10.1186/gb-2007-8-8-r156.
93. Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: Connectivity Rather Than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol* **28**, 1481–1489 (2011).
94. Acar Kirit, H., Lagator, M. & Bollback, J. P. Experimental determination of evolutionary barriers to horizontal gene transfer. *BMC Microbiol* **20**, 1–13 (2020).
95. Igler, C. *et al.* Limited specificity of molecular interactions incurs an environment-dependent fitness cost in bacteria. *bioRxiv* 2021.10.20.465141 (2021) doi:10.1101/2021.10.20.465141.
96. Haldane, A., Manhart, M. & Morozov, A. V. Biophysical Fitness Landscapes for Transcription Factor Binding Sites. *PLoS Comput Biol* **10**, 36–38 (2014).
97. Mustonen, V., Kinney, J., Callan, C. G. & Lassig, M. Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences* (2008) doi:10.1073/pnas.0805909105.
98. Tuğrul, M., Paixão, T., Barton, N. H. & Tkačik, G. Dynamics of Transcription Factor Binding Site Evolution. *PLoS Genet* **11**, 1–28 (2015).
99. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* **45**, 723–729 (2013).
100. Mustonen, V. & Lassig, M. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences* **102**, 15936–15941 (2005).
101. Crocker, J., Preger-Ben Noon, E. & Stern, D. L. *The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution*. *Current Topics in Developmental Biology* vol. 117 (Elsevier Inc., 2016).
102. Kurafeiski, J. D., Pinto, P. & Bornberg-Bauer, E. Evolutionary Potential of Cis-Regulatory Mutations to Cause Rapid Changes in Transcription Factor Binding. *Genome Biol Evol* **11**, 406–414 (2019).
103. Schulz, S. *et al.* Elucidation of Sigma Factor-Associated Networks in Pseudomonas aeruginosa Reveals a Modular Architecture with Limited and Function-Specific Crosstalk. *PLoS Pathog* **11**, 1–21 (2015).
104. Martin, A. *et al.* Multiple recent co-options of Optix associated with novel traits in adaptive butterfly wing radiations. *Evodevo* **5**, 7 (2014).
105. Cvekl, A. *et al.* Evolutionary origins of Pax6 control of crystallin genes. *Genome Biol Evol* (2017) doi:10.1093/gbe/evx153.
106. Shepherd, M. J., Pierce, A. P. & Taylor, T. B. Evolutionary innovation through transcription factor rewiring in microbes is shaped by levels of transcription factor activity, expression, and existing connectivit. *PLoS Biol* **21**, e3002348 (2023).

107. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).
108. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic Gene Expression in a Single Cell. *Science (1979)* **297**, 1183–6 (2007).
109. Sharon, E. *et al.* Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res* **24**, 1698–1706 (2014).
110. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521–530 (2012).
111. Rudge, T. J. *et al.* Characterization of Intrinsic Properties of Promoters. *ACS Synth Biol* **5**, 89–98 (2016).
112. Beal, J. *et al.* Quantification of bacterial fluorescence using independent calibrants. *PLoS One* **13**, e0199432 (2018).
113. Beal, J. *et al.* Reproducibility of fluorescent expression from engineered biological constructs in E. coli. *PLoS One* **11**, (2016).
114. Beal, J., Haddock-Angelli, T., Farny, N. & Rettberg, R. Time to Get Serious about Measurement in Synthetic Biology. *Trends in Biotechnology* (2018) doi:10.1016/j.tibtech.2018.05.003.
115. Trippe, B. L. *et al.* Randomized gates eliminate bias in sort-seq assays. *Protein Science* **31**, e4401 (2022).
116. Peterman, N. & Levine, E. Sort-seq under the hood: Implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 1–17 (2016).
117. Gilliot, P.-A. & Gorochowski, T. E. Effective design and inference for cell sorting and sequencing based massively parallel reporter assays. *bioRxiv* 2022.11.07.515414 (2022) doi:10.1101/2022.11.07.515414.
118. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology 2019 38:1* **38**, 56–65 (2019).
119. Belliveau, N. M. *et al.* A Systematic and Scalable Approach for Dissecting the Molecular Mechanisms of Transcriptional Regulation in Bacteria. *Biophys J* **114**, 151a (2018).
120. Lagator, M. *et al.* Predicting bacterial promoter function and evolution from random sequences. *Elife* **11**, (2022).

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 and the UZH flow cytometry facility for their technical support. We are especially thankful to Diego Pesce for his assistance in establishing experimental protocols, and to Andrei Papkou for his guidance with computational analysis and for engaging in theoretical discussions.

Author contributions: C.A.W. and A.W conceived the study and designed experiments. C.A.W. carried out experiments. C.A.W. wrote computer code to carry out bioinformatic work and data analysis. A.W. provided analytical guidance in data 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.

Competing interests: The authors declare no competing interest

Corresponding authors: Correspondence to andreas.wagner@ieu.uzh.ch

Figure legends:

Figure 1. Experimental workflow. a-b. The reporter system and its regulatory modules. We redesigned a previously developed plasmid[31] (**Supplementary Methods 2-3**) that encodes a chloramphenicol resistance gene as a selective marker (CmR, grey, right-oriented arrow) and features a broad-host low-copy replication origin (pBBR1, grey, left-oriented arrow). This plasmid harbors three modules interconnected through regulatory interactions. The first module comprises the constitutively expressed *tetR* gene (purple). The TetR regulator controls the second regulatory module (red box) through the *plTetO1* promoter[32]. This second module encompasses a bicistronic operon, comprising both the TF of interest (red) and a reporter *mScarlet-I* gene[33] (red). In the third module (green box), the TF interacts with its cognate TFBS or a variant from the appropriate library. This interaction results in the transcriptional repression of the fluorescent reporter gene *sfgfp*[34] through steric hindrance, blocking RNA polymerase activity. The addition of anhydrotetracycline (Atc) alleviates repression by TetR, prompting the expression of the TF of interest, which leads to repression of *sfgfp*. To safeguard against the unintended effects of 5'- untranslated regions (UTRs) created by some TFBS library variants on the expression of the *sfgfp* mRNA, we used the RiboJ[35,36] transcriptional insulator (hairpin structure, in grey). Ribosome binding sites (RBSs) are highlighted as grey semi-circles. **c. Basic reporter dynamics.** As the TetR inducer (Atc) levels increase, more TetR regulator is sequestered from its binding site, allowing the expression of the TF and the mScarlet-I reporter (red). Increasing TF concentration causes a decrease in green fluorescence, because the TF causes increasing repression of the sfGFP reporter. **d. Library generation.** Alignment of two strong-affinity ("wild-type") TFBSs for individual TFs, indicating positions (dots) at which the two aligned sites differ from each other. We created a binding site library by varying the nucleotides at all nucleotide positions that differ between the two sites. The library comprises 2^n sequences, where n is the number of positions at which the sites differ (See **Supplementary Methods 4, Supplementary Table S3** and **Supplementary Figure S3** for library details). **e. Transformation of the reporter system harboring the library into two distinct bacterial strains.** One strain expresses TF1 (+TF1) but not TF2 (ΔTF2), and the other TF2 (+TF2) but not TF1(ΔTF1). We then performed fluorescence-activated cell sorting (FACS) independently for each strain, separating cells (green rectangles) into specific bins according to their fluorescence. High-throughput sequencing of the library variants from each bin reveals the mutational differences in each variant relative to the reference TFBS, indicated as red and blue

bars inside grey cylinders (TFBS), corresponding to data from the two TF-producing strains. The strength of regulation (binding) conveyed by each binding site is measured from the site's read count across bins (**Supplementary Methods 7.2**). **f. Mapping an Exaptation landscape.** Each TFBS is characterized by two regulation strengths, one for each TF. The regulation strength of each genotype defines an adaptive landscape for each TF (shown as the central-colored network, with the color gradient representing the regulation strengths for TF1, red, and TF2, blue). These can be interpreted as two entangled adaptive landscapes (red and blue above and below the network), where a single sequence is mapped to two phenotypes.

Figure 2 a-c. Specific plasmid-strain combinations ensure the presence of only one of two TFs during our measurements. Each table schematically represents the intracellular presence of TFs for combinations of a reporter plasmid encoding one of the TFs (rows) transformed into a mutant *E. coli* strain lacking the gene encoding one of the TFs (columns). Pure colors represent the presence of a single TF in the cell (CRP: red; Fis: green and IHF: purple), and color gradients represent the presence of both TFs. We avoided plasmid-strain combinations in which both TFs were present in the cell, as such combinations did not allow us to unambiguously measure the regulation conveyed by one TF's binding to each of the TFBSs in an exaptation landscape. The plasmid-strain combinations we used are highlighted with thick black borders. **d-f. Distribution of regulation strengths for three TFBS libraries.** Each panel shows data from a different TF1-TF2 library (d. CRP-Fis; e. CRP-IHF; f. Fis-IHF). Each circle represents a single TFBS variant in a landscape. The horizontal and vertical axes represent the regulation strengths of each variant in the respective landscape to the TF labelled on the axes. Pure colors represent the regulation strengths of each TF (CRP: Red; Fis: Green: IHF, blue), and intermediate colors represent binding by two TFs. The histograms on the top and right of each plot show the distribution of regulation strengths for each TF.

Figure 3. Exaptation landscapes are smooth with many intermediates subject to transcriptional crosstalk. Panels a-c depict exaptation landscapes as networks whose nodes (circles) represent binding site variants connected through single nucleotide changing mutations (straight lines). The x-axis of each panel shows the genetic distance (in number of nucleotide differences) from each TFBS variant to one of the wild-type sequences (at x = 0). Grey squares indicate WT sequences for each TF, which are peaks in their respective landscapes. **a. The CRP-Fis exaptation landscape.** Both panels represent the CRP-Fis exaptation landscape (N = 256) with nodes ordered by their distance to the CRP wild-type sequence. Colors represent regulation strengths of each TFBS variant to CRP (top panel, red heatmap) and to Fis (bottom panel, green heatmap). B**. The CRP-IHF exaptation landscape.** Both panels represent the CRP-IHF exaptation landscape (N = 128) with nodes ordered by their distance to the CRP wild-type sequence. Colors represent regulation strengths of each TFBS variant to CRP (top panel, red heatmap) and to IHF (bottom panel, purple heatmap). **c. The Fis-IHF exaptation landscape.** Both panels represent the Fis-IHF exaptation landscape (N = 128) with nodes ordered by their distance to the Fis wild-type sequence. Colors represent Fis (top panel, green heatmap) and IHF (bottom panel, purple heatmap) regulation strengths for each sequence. Light and dark colors in panels a-c represent weak and strong binding, respectively. Note that all regulation strengths in panels a-c are normalized by the regulation strength of the wild-type sequence. **d-f. Crosstalk levels for binding sites in each exaptation landscape.** Each panel shows, for all TFBSs in an exaptation landscape, the ratio of binding to TF1 and TF2 (y-axis) as a function of the TFBSs genetic distance from the wild-type sequence (x-axis), which is a strong TFBS. Each circle symbolizes one genotype. Grey circles connect genotypes that differ in a single nucleotide. Boxplots indicate the distribution of crosstalk levels for each mutational distance. Each box covers the

range between the first and third quartiles (interquartile range, IQR). The horizontal line within the box represents the median value, and whiskers span 1.5 times the IQR. Heatmap colors reflect the regulation strength ratio for each TFBS. That is, sequences binding predominantly to CRP, Fis or IHF are displayed in red, green, and purple, respectively. Intermediate colors indicate binding at intermediate strength to two TFs. Note that crosstalk levels are normalized by the maximum crosstalk value measured, resulting in a scale that ranges from 0 (no crosstalk) to 1 (maximum crosstalk). We generated networks with the *Gephi*[38] software using a combination of both the *3DForceAtlas* and *Network3DSplitter* algorithms.

Figure 4. Rapid exaptation on the CRP-Fis landscape. All data presented is based on a population of 10^8 individuals. Data for each plot is based on 10^5 simulated adaptive walks, i.e., 10^4 adaptive walks for each of the 10 starting genotypes with the highest regulation strength for the opposite TF. **a-b. Vector field illustrations of adaptive walks in genotype space.** The illustrations show the direction of adaptive walks in genotype space when strong binding to a specific TF (CRP or Fis) is favored. Each panel represents a single TF landscape, with the x and y axes representing coordinates in genotype space. Each vector indicates the direction of an adaptive walk starting from a distinct genotype and ending at the colored circle peak ("attractor"), which signifies strong regulation by the TF. **a. Scenario 1: Adaptive trajectories favoring strong binding to CRP.** In the CRP landscape, adaptive walks favor strong binding to CRP, indicated by the orange circle (CRP peak). **b. Scenario 2: Adaptive trajectories favoring strong binding to Fis.** In the Fis landscape, adaptive walks favor strong binding to Fis, indicated by the green circle (Fis Peak). **c-d. Adaptive walks readily attain peaks.** Each plot represents how regulation strengths (vertical axis) for both TFs (CRP in orange and Fis in green) change along the transversed path length (horizontal axis) from the initial sequence to the attained peak during an adaptive walk. Curves represent averages over all 10^5 adaptive walks, and shading indicates one standard deviation of regulation strengths. **c. Adaptive walks attain the CRP peak.** The data is based on 10^4 walks starting from each of the 10 strongest binding sites for Fis, favoring increasing CRP binding. **d. Adaptive walks attain the Fis peak.** Like panel (c), but for adaptive walks starting from the 10 strongest binding sites for CRP, favoring increasing Fis binding. **e-f. Evolutionary paths to a peak are not much longer than the minimal genetic distances.** Each composite histogram shows the distribution of path lengths (green) for all pairs of starting (non-peak) and ending (peak) genotypes during 10^5 adaptive walks. Blue histograms represent the distribution of genetic distances for each pair of non-peak and peak genotypes. The green and blue horizontal lines represent average path lengths and mutational distance, which are indistinguishable. We used a t-test of the null hypothesis that the means of these distance distributions are statistically indistinguishable. **e. Distribution of accessible path lengths and genetic distances to the CRP peak.** We observed no significant difference between mean genetic distance and path length (mean ± s.d.: 6.6 ± 0.7 for both genetic distance and path length; Welch Two Sample t-test: $t = 31.973$, $df = 196{,}700$, p-value = 0.7, $N_1 = 10^5$ genetic distances, $N_2 = 10^5$ path lengths). **f. Distribution of accessible path lengths and genetic distances to the Fis peak.** We observed no significant difference between mean genetic distance and path length (mean ± s.d.: 6.9 ± 0.8; Welch Two Sample t-test: $t = 35.113$, $df = 196{,}700$, p-value = 0.82, $N_1 = 10^5$ genetic distances, $N_2 = 10^5$ path lengths).

Figure 5. Transcriptional crosstalk and its consequences for evolutionary dynamics. a. PWM scores for experimentally characterized TFBSs from the *E. coli K-12* genome suggest ample crosstalk between TFs CRP, Fis, and IHF. The heatmap is based on 756 TFBS sequences in the *E. coli K-12* genome, as curated in the RegulonDB[67] database. Each row corresponds to one TFBS. Specifically, the dataset includes 370 TFBSs for CRP (orange in left-most column), 267 for Fis (green), and 119 for IHF (purple). Each of the remaining columns

quantifies the predicted binding strength (PWM score) for each of the 756 TFBSs with respect to the transcription factors CRP, Fis, and IHF, as indicated below each column. We normalized predicted binding strength by the maximum possible score for each TF, resulting in a scale from 0 to 1, as indicated by the color gradient in the heatmap color legend. We clustered the rows of the heatmap using a combination of Euclidean distance and hierarchical clustering with complete-linkage to reveal patterns of similarity in predicted binding strengths among the sequences. **b-d. Association between predicted binding strengths for CRP, Fis and IHF and their genomic *E. coli* TFBS.** We computed for each pair of TFs the statistical association between predicted binding strength (CRP-Fis, CRP-IHF, Fis-IHF). In the scatter plots, red circles represent TFBSs whose average predicted binding strength for both TFs is higher than 0.5, indicating intermediate to high crosstalk. Grey circles represent TFBSs with an average predicted binding strength below this threshold, indicating low crosstalk. The percentage of sequences with intermediate to high crosstalk (red circles) and low crosstalk (grey circles) is shown at the top of the plots in red and gray. The blue line is a linear regression line (blue shading: 95% confidence interval; R: Pearson correlation coefficient.). **b. Association between CRP and IHF.** No significant association between CRP and Fis predicted binding strengths ($R = -0.01$ and t-statistic $p = 0.71$), which is due to the clear distinction between sparsely scattered points and the straight line involving some sequences with very similar predicted binding strength values. **c. Association between CRP and IHF.** $R = -0.56$ and t-statistic p-value $< 2 \times 10^{-6}$. **d. Association between Fis and IHF.** $R = -0.24$ and t-statistic p-value $< 2 \times 10^{-6}$. **e-g. Rapid evolution of crosstalk in the CRP-Fis landscape. e. Schematic of evolution towards TFBSs with high crosstalk.** The vector field illustrates the direction of adaptive walks in genotype space when strong binding towards a CRP and Fis is favored. Each vector indicates the direction of an adaptive walk starting from a distinct genotype and ending at a peak ("attractor", colored circle) strongly binding both TFs. **f. Adaptive walks readily attain TFBSs binding to both CRP and Fis with intermediate strengths.** We performed 10^4 adaptive walks starting from each of the 256 TFBS variants in the CRP-Fis landscape, with selection favoring strong binding for both CRP and IHF. The panels show the regulation strength (vertical axis) for CRP (orange) and Fis (green) plotted against the transversed path length of an evolving sequence to the attained peak (horizontal axis). Curves represent averages over all 256×10^4 adaptive walks and shading indicates one standard deviation of regulation strengths. **g. The distribution of accessible path lengths and genetic distances to peaks.** The green and blue horizontal lines show the average length of traversed paths during 256×10^4 walks between pairs of non-peaks and peaks, and the respective mutational distances between such pairs, respectively. The means of the distributions are statistically indistinguishable (3.31 ± 1, mean ± s.d., for both distance metrics, Welch Two Sample t-test $t = -5.5782$, $df = 2029535$, p-value = 0.7, $N_1 = 256 \times 10^4$, $N_2 = 256 \times 10^4$).