HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2021 December 29.

Published in final edited form as:

Cell Rep. 2021 August 24; 36(8): 109590. doi:10.1016/j.celrep.2021.109590.

Systematic dissection of σ^{70} sequence diversity and function in bacteria

Jimin Park[1,2,*], **Harris H. Wang**[1,3,4,*]

[1]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA

[2]Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University Irving Medical Center, New York, NY, USA

[3]Department of Pathology and Cell Biology, Columbia University Irving Medical Center, New York, NY, USA

[4]Lead contact

SUMMARY

Primary σ^{70} factors are key conserved bacterial regulatory proteins that interact with regulatory DNA to control gene expression. It is, however, poorly understood whether σ^{70} sequence diversity in different bacteria reflects functional differences. Here, we employ comparative and functional genomics to explore the sequence and function relationship of primary σ^{70}. Using multiplex automated genome engineering and deep sequencing (MAGE-seq), we generate a saturation mutagenesis library and high-resolution fitness map of *E. coli* σ^{70} in domains 2–4. Mapping natural σ^{70} sequence diversity to the *E. coli* σ^{70} fitness landscape reveals significant predicted fitness deficits across σ^{70} orthologs. Interestingly, these predicted deficits are larger than observed fitness changes for 15 σ^{70} orthologs introduced into *E. coli*. Finally, we use a multiplexed transcriptional reporter assay and RNA sequencing (RNA-seq) to explore functional differences of several σ^{70} orthologs. This work provides an in-depth analysis of σ^{70} sequence and function to improve efforts to understand the evolution and engineering potential of this global regulator.

In brief

Through comparative and functional genomics, Park and Wang dissect the sequence and function relationship of the bacterial σ^{70} factor. MAGE-seq generates a saturation mutagenesis library and a high-resolution fitness map of *E. coli* σ^{70}. Replacement of endogenous *E.coli* σ^{70} with natural orthologs elicits transcriptional changes.

*Correspondence: jiminpark66@gmail.com (J.P.), hw2429@columbia.edu (H.H.W.).

AUTHOR CONTRIBUTIONS

J.P. and H.H.W. conceived the study. J.P. performed all experiments and data analysis with supervision from H.H.W. Both authors wrote the paper together.

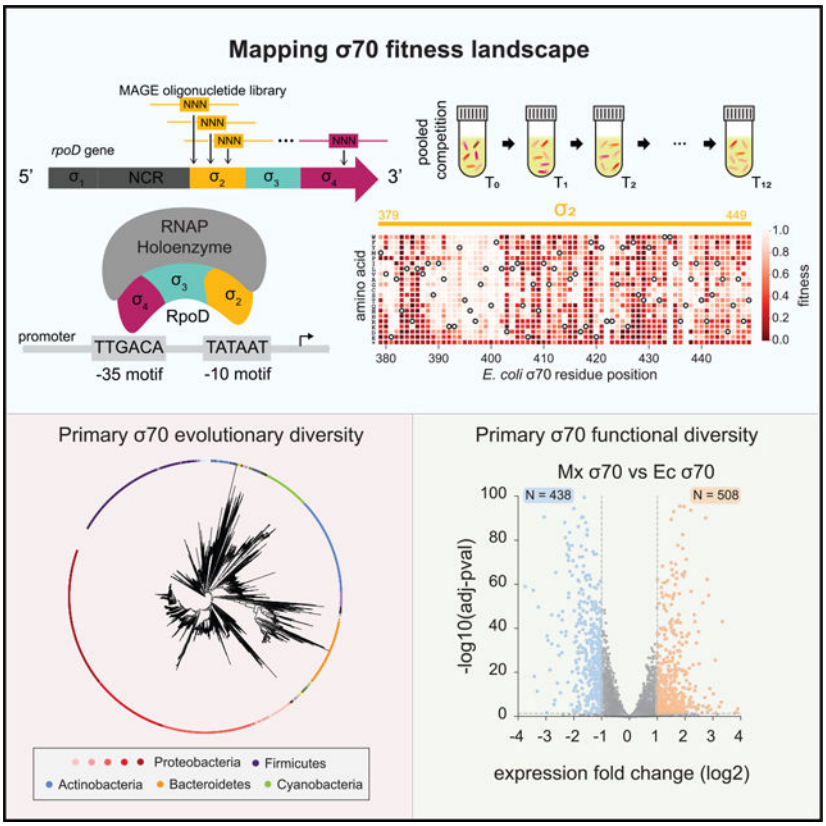
SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109590.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Graphical Abstract



INTRODUCTION

Bacterial gene expression is coordinated through interactions between *cis*-regulatory DNA sequences and *trans*-regulatory proteins to facilitate cell growth, adaptation, and response to external stimuli (Browning and Busby, 2016; Phillips et al., 2019). During transcription, regulatory proteins recognize sequence motifs in the 5′ regulatory regions (e.g., promoters) upstream of protein-encoding genes to control their expression. A key class of regulatory proteins in bacteria is the sigma factors that control many essential functions in the cell. To coordinate gene expression, sigma factors interact with an RNA polymerase (RNAP) core enzyme (consisting of $\alpha_2\beta\beta'\omega$ subunits) to form the RNAP holoenzyme and directs the complex to promoter regions by recognizing specific regulatory signatures (Feklístov et al., 2014; Paget and Helmann, 2003). There, sigma factors unwind the DNA duplex and facilitate transcription initiation. Sigma factors are classified into either σ^{70} or σ^{54} protein families. The σ^{70} protein family contains primary (group 1) and alternative (group 2–4) sigma factors and recognizes the −10/−35 promoter motifs. The σ^{54} protein family has a different recognition domain that recognizes the −12/−24 promoter motifs and is functionally, structurally, and evolutionarily distinct from σ^{70} (Merrick, 1993). In most bacteria, the primary σ^{70} protein is known as *rpoD* or *sigA* and controls the expression of the largest fraction of genes in the cell. The remaining alternative σ^{70} factors regulate more targeted cellular functions such as flagellar proteins or responses to environmental

stressors (Feklístov et al., 2014; Paget and Helmann, 2003). In *E. coli*, the primary σ^{70} factor accounts for 60%–95% of all sigma factors present in the cell during exponential growth and binds to >50% of all sigma factor binding sites across the genome (Gama-Castro et al., 2016; Grigorova et al., 2006; Ishihama, 2000).

The primary σ^{70} factor encodes four conserved protein domains (Feklístov et al., 2014; Paget and Helmann, 2003). Domain 1 plays an inhibitory role, preventing free σ^{70} proteins from binding to DNA while not in complex with the rest of the RNAP subunits. Domain 2 recognizes the −10 promoter motif and facilitates unwinding of the DNA duplex for transcription initiation, while domain 3 interacts with the extended −10 promoter motif. Domain 4 mediates recognition of the −35 promoter motif. Since primary σ^{70} is highly conserved across bacteria, promoter motifs similar to the canonical *E.coli* σ^{70} motif have been reported from diverse bacterial species (Bottacini et al., 2017; Domínguez-Cuevas and Marqués, 2004; Jeong et al., 2016; Moran et al., 1982; Rosinski-Chupin et al., 2015; Sharma et al., 2010). However, σ^{70}-associated regulatory sequences can be quite diverse even within a single genome such that any single sequence motif will not necessarily predict transcriptional output (Urtecho et al., 2019).

Changes to primary σ^{70} factors have been shown to alter the cellular transcriptome. Mutants of the *E. coli* σ^{70} generated through laboratory evolution exhibited genome-wide transcriptional changes that yielded novel cellular phenotypes such as improved tolerance to environmental stresses (Alper and Stephanopoulos, 2007). Furthermore, many point mutations in σ^{70} domain 2 or 4 have been generated and characterized in the past several decades (Gardella et al., 1989; Siegele et al., 1989; Waldburger et al., 1990). These studies highlighted that single mutations in functional domains alone were sufficient to elicit altered gene expression patterns from various model regulatory sequences. Furthermore, heterologous expression of a σ^{70} ortholog in *E. coli* also resulted in recognition of non-native regulatory DNA, highlighting the flexibility of many σ^{70} factor orthologs to interact with RNAP to recognize non-native transcriptional signatures (Gaida et al., 2015; Tomko and Dunlop, 2017). While these results suggest that σ^{70} can be evolved or engineered to tune transcription in a variety of ways, the impact of specific σ^{70} mutations on the global transcriptome is still not fully understood.

Here, we performed systematic computational and high-throughput experimental studies to profile the sequence-function relationship of σ^{70} and its impact on host gene expression and fitness. Using comparative genomics, we first explored the evolutionary diversity of σ^{70} across the tree of life to understand the functional conservation of key residues and domains of this global regulator. We then employed deep mutational scanning to systematically dissect the impact of individual residue mutations on σ^{70} function in *E. coli* to build out the most comprehensive experimentally generated fitness landscape of a bacterial σ^{70} to date. Variants from these fitness measurements could be mapped to natural σ^{70} orthologs to assess functional selection during σ^{70} evolution. Replacement of the endogenous *E. coli* σ^{70} with natural orthologs revealed large-scale transcriptome rewiring that could be further probed using a multiplexed transcriptional reporter assay to dissect determinants of transcription. These results offer a high-resolution map of the evolutionary landscape of a bacterial primary σ^{70} factor and its transcriptomic function during evolution.

RESULTS

Evolutionary diversity of σ^{70} across bacteria

We first sought to systematically profile bacterial σ^{70} diversity by mining group 1 σ^{70} orthologs of *E. coli* RpoD from the KEGG ortholog database (Kanehisa et al., 2016; see STAR Methods), which yielded ~4,700 sequence variants from mostly Proteobacteria (~42.6%), Firmicutes (~15%), Actinobacteria (~15%), and Bacteroidetes (~8%). A phylogenetic tree based on multiple sequence alignment (MSA) of these σ^{70} orthologs closely recapitulated the phylum- and class-level organizations of their respective genomes of origin (Figure S1A). In addition, the phylogenetic distances between *E. coli* RpoD and σ^{70} orthologs correlated well with the 16S divergence of *E. coli* to their corresponding bacteria, as expected for a conserved protein that has also been used as a phylogenetic marker (Gruber and Bryant, 1997; Figure S1B). At the residue level, the highest amino acid conservation (as measured by Jensen-Shannon divergence; Capra and Singh, 2007) was observed in RpoD throughout domains 2–4, but not in domain 1, with the exception of domain 1.2 (Figure 1A). These evolutionary conservation patterns reflect key functional regions of RpoD, with domain 2 (residues 379–449) binding to the −10 promoter motif and unwinding the DNA duplex, domain 3 (residues 458–535) interacting with the extended −10 motif, and domain 4 (residues 547–600) binding to the −35 motif (Feklístov et al., 2014; Paget and Helmann, 2003). Next, we used the *E. coli* RpoD, one of the most extensively studied primary σ^{70} orthologs (others being SigA from *Thermus aquaticus*; Feklistov and Darst, 2011; Murakami and Darst, 2003; and *Mycobacterium tuberculosis*; Manganelli et al., 2004; Rodrigue et al., 2006) as a “reference sequence” to study σ^{70} sequence diversity, in part also because of the possibility to experimentally alter *E. coli* RpoD by genome engineering. Each of the four σ^{70} domains in diverse RpoD orthologs had varying amounts of residue differences (i.e., substitutions) from *E. coli* RpoD (Figure 1B). Domains 2 and 4 had the lowest fraction of residue differences, which reflected strong evolutionary conservation in these domains. Furthermore, the degree of residue differences in each domain was well correlated with the 16S phylogenetic distances of σ^{70} orthologs to *E. coli* as well as the full-length σ^{70} phylogenetic distances to *E. coli* σ^{70} (Figures S2A and S2B). Given the functional importance and high evolutionary conservation of domains 2–4, we limited our subsequent analyses and studies to these domains (matching to amino acid positions 379–613 of *E. coli* RpoD) and used a subset of 2,833 unique RpoD variants.

The distribution of residue differences in domain 2 of σ^{70} orthologs exhibited three distinct peaks centered on zero, ~15, or ~33 substitutions, which suggests three major branches of sequence variants. Comparatively, a more uniform distribution of substitutions was observed in domains 3 and 4, with peaks near the median residue difference counts, corresponding to ~35 and ~20 substitutions, respectively. Next, we sought to further analyze domain-level diversity with a more unbiased analysis method that does not depend on a starting reference sequence variant. By clustering domain sequences by similarity, we observed that in domain 2 majority of orthologs (1,575 out of 2,833) could be captured by four major sequence clusters mapping to Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes (Figure 1C). As expected, domains 3 and 4, on the other hand, required additional sequence clusters to capture diversity of similar number of orthologs as domain 2. These analyses suggest that

molecular evolution in domains 3 and 4 produced a more continuous spectrum of sequence variations than that in domain 2. Lastly, the number of unique amino acids observed at each residue position across domains 2–4 also showed that domain 2 residues had access to fewer unique amino acids compared to residues in domains 3 or 4 (Figure S2C).

To better understand inter-domain σ^{70} evolution, we performed pairwise comparisons of the substitution ratios between domains (Figure 1D). Substitution ratios were calculated by dividing substitution counts in each domain by the length of its specific σ^{70} domain; this normalization allowed us to compare the degree of substitution between domains. Higher substitution ratios were observed in domain 3 than in domain 4, and higher ratios were observed in domain 4 than in domain 2. This observation suggests that, compared to *E. coli* RpoD as reference, phylogenetically closest orthologs have only domain 3 diversity, more distant orthologs have greater domain 3 and 4 diversity, and the most phylogenetically distant orthologs exhibit diversity in all three domains. As functionally important residues are more likely to have fewer viable substitutions available, the observed domain diversity pattern may reflect functional ordering of σ^{70} domains, with domain 2 being more important than domain 4 and domain 4 being more important than domain 3. Lastly, to map and compare domain divergence against phylogenetic diversity, we compared substitution ratios of each domain stratified by phyla (Figure S3A) and found that domain-specific relationships between substitution ratios and phyla were largely similar among all three domains. These results highlight important domain-specific patterns of sequence diversity reflecting evolutionary diversification of σ^{70}.

Next, we explored whether various genomic signatures could help explain the observed patterns of σ^{70} diversity. While genomic GC content has a strong influence on gene expression (Johns et al., 2018), we did not observe any relationship between GC content and RpoD residue substitution patterns (Figure S3B). We then analyzed the *cis*-regulatory regions (i.e., upstream of start codons), which are bound by σ^{70} during transcription initiation, to determine if sequence motifs identified in the *cis*-regions correlated with σ^{70} diversity. Sequences upstream of putative σ^{70}-regulated genes in diverse bacteria were used to predict σ^{70}-associated regulatory sequence motifs using BioProspector (Liu et al., 2001; see STAR Methods). Hierarchical clustering of the motif score correlations showed that as expected, many bacteria used similar canonical −10 and −35 σ^{70} motifs, while some had more divergent motifs (Figure S3C). Interestingly, we observed a weak but significant inverse relationship between degree of similarity to the *E. coli* −10/−35 motif and degree of divergence from *E. coli* RpoD (Figure S3D). At the domain level, domain 4 showed the strongest relationship between RpoD conservation and motif similarity, suggesting that this domain co-evolved with the regulatory DNA more than the other domains. Together, our results so far provide insights into global σ^{70} evolution from a comparative genomics lens that could be further contextualized with systematic experimental data to reveal relationships between σ^{70} sequence and function.

Systematic dissection of the *E. coli* σ^{70} fitness landscape

To better understand the functional differences between σ^{70} variants, we sought to systematically profile its sequence-function relationship by high-throughput mutagenesis

and phenotypic measurements. Because σ^{70} initiates a majority of cellular transcripts during exponential growth, cellular fitness is significantly affected by any σ^{70} mutations that alter its function as a global transcription factor. To systematically interrogate the impact of σ^{70} mutations on cellular fitness and gene expression, we used the *E. coli* RpoD as a model and targeted the single-copy *rpoD* gene in the *E. coli* genome for saturation mutagenesis. We previously showed that high-efficiency oligo-recombineering enables direct genomic mutagenesis and fitness measurements of essential genes in a pooled format (Kelsic et al., 2016). Accordingly, we used MAGE-seq (multiplex automated genome engineering and deep sequencing; Wang et al., 2009) to generate a comprehensive mutagenesis library of RpoD along domains 2–4, the most functionally interesting regions (Figures S4A and S4B). 236 MAGE oligos were designed to each target a single-residue position tiled across domains 2–4 (residues 379–613 and stop codon) of RpoD with degenerate NNN sequences to create all 64 codon variants (see STAR Methods). This oligo library yielded a total of ~15,000 nucleotide variants or ~4,700 amino acid variants. After six rounds of MAGE in *E. coli* using this oligo library, the *rpoD* gene was amplified from the mutagenized cell population by PCR and analyzed by deep sequencing, which showed that ~18% of the population carried single-codon *rpoD* mutations (Figure S4C). After correcting for sequencing errors (STAR Methods; Figure S4D), 14,576 out of 14,868 total possible variants could be detected in the population, representing a >98% coverage of the mutational sequence space.

To determine the fitness of individual *rpoD* variants, we performed pooled growth competition on the mutagenized population over time. Competition experiments were carried out in a turbidostat with at least 10^8 cells under exponential growth to prevent population bottlenecks and maintain a constant growth selection pressure. The cell population was sampled at regular intervals, and *rpoD* variant frequencies were assessed by deep sequencing. The relative fitness of each mutant was determined by fitting the relative change in variant frequency over time to a log-linear regression compared to wild-type (WT) *rpoD* (Figure S4B). A fitness of 1 meant that a mutant had an equivalent growth rate as WT RpoD, while a fitness of 0 represented no measurable growth and subsequent loss from the population at the turbidostat dilution rate. These growth measurements yielded a near-comprehensive fitness landscape of all single-codon variants of RpoD (Figures 2A, S5A, and S5B).

The RpoD fitness map showed both expected and novel features of the protein. As anticipated, premature stop codons had a low mean fitness of 0.23, with a standard deviation of 0.25 (Figure S5C). The non-zero fitness of premature stop codons could be attributed to experimental noise at low-growth regimes and residual RpoD proteins that may have contributed to some background growth. Interestingly, premature stop codons were tolerated in the last three residues, suggesting that RpoD could support C-terminal truncations of the last three residues without any negative functional impact. On the other hand, sense mutation of the WT *rpoD* stop codon showed reduced fitness, suggesting that a C-terminal translational run-on is not well tolerated (i.e., the next in frame downstream stop codon is 49 residues away). As expected, synonymous mutations showed little fitness impact, while non-synonymous mutations had a wide range of fitness effects (Figure S5D). Notably, proline substitutions, which increase conformational rigidity and often significantly change protein

secondary structure, exhibited low fitness throughout RpoD. Using principal-component (PC) analysis, we further assessed the global biochemical determinants of RpoD and found that the first four PCs could explain ~85% of the variance in the fitness data (Figure S6). These PCs matched key amino acid biochemical properties including free energy (PC1), hydrophobicity (PC2), steric hinderance or size (PC3), and helices (PC4).

To better assess the RpoD fitness landscape, we analyzed its distribution of fitness effects (DFEs) (Figure 2B). The RpoD DFE is bimodally distributed, with a narrow fitness peak centered near 1 (i.e., neutral mutations) and a wide fitness peak centered at ~0.25 (i.e., detrimental mutations) (Figure S5B). From other systematic mutagenesis studies (Firnberg et al., 2014; Jacquier et al., 2013; Kelsic et al., 2016; Konaté et al., 2019; Sarkisyan et al., 2016; de Visser and Krug, 2014), bimodal DFEs are commonly observed in fitness landscapes of many proteins that are essential for cellular function, confer antibiotic resistance, or produce fluorescence. Using a fitness threshold of 0.95 (i.e., 3 standard deviations below the mean fitness of synonymous mutations) to separate between neutral (≥0.95) and deleterious (< 0.95) mutations, we find that 52.9% of RpoD mutants were deleterious. In comparison, 48% and 38% of mutations in essential proteins IF1 and DHFR, respectively, were deleterious based on similar MAGE-seq fitness measurements (Kelsic et al., 2016; Konaté et al., 2019). At the domain level, 70.9% of domain 3 mutations were near neutral (≥0.95) compared to 22.4% and 36.6% for domain 2 and domain 4, respectively (Figure 2B). Therefore, the mutational paths that do not yield significant negative fitness effects are notably more restricted in domains 2 and 4 than in domain 3.

The mutational fitness patterns of *E. coli* RpoD also matched natural RpoD evolution. Generally, neutral residues are not evolutionarily conserved, while functionally important residues are highly conserved. As expected, plotting the mean fitness of all amino acid substitutions against the evolutionary conservation for each residue position revealed an inverse correlation (Figure 2C). To determine the relationship between mean fitness of substitutions and conservation between RpoD domains, we compared the linear regression slopes of non-neutral (mean fitness < 0.95) residues in each domain. Upon further stratification by domain, we observed that the slopes between fitness and conservation were different between domains. Domain 3 had the flattest slope (−0.65 ± 0.22) followed by domain 2 (−1.26 ± 0.45) and then domain 4, which had the steepest slope (−2.12 ± 0.34). These slopes suggest that fitness impacts differed between domains for residues with similar degrees of conservation. As such, different selective forces may be driving the evolution of each RpoD domain separately.

We next sought to better contextualize the RpoD fitness landscape at a residue position level resolution by leveraging the wealth of structure-function literature available for primary σ^{70}. First, to facilitate straightforward comparisons between literature-derived functional residues and fitness landscape derived functional residues, we applied hierarchical clustering to the fitness landscape, yielding three discrete types of residue positions, with 37 positions (15% of the protein) being highly deleterious (almost all substitutions exhibit fitness costs), 76 positions (33%) being variably deleterious (some amino acid substitutions exhibit fitness costs while others are neutral), and 118 positions (51%) being near neutral (most substitutions are neutral) (Figure S7A). From the literature, a set of 63 residue positions

that been reported to be important for RpoD function was compiled (Bae et al., 2015; Campbell et al., 2002; Feklistov and Darst, 2011; Feng et al., 2016; Fenton et al., 2000; Hook-Barnard and Hinton, 2007; Panaghie et al., 2000; Zhang et al., 2012; Figure S7B). One of the core RpoD function is −10 motif recognition, which is mediated through residues in domain 2. Specifically, the stretch of residues between positions 383 and 389, which is known to mediate interactions with the −7 position of the −10 motif, and positions 414–429, which are reported to mediate interactions with the −12 and −11 position, were regions associated with high fitness defects. DNA duplex unwinding at the −10 motif is mediated through the two tryptophan residues at 433 and 434; mutagenesis of residue 433 yielded highly deleterious phenotypes (no variants with mutations at position 434 were recovered). Furthermore, the extended −10 motif, is another *cis*-regulatory motif that can regulate transcription, especially for promoters that lack a −35 motif (Paget and Helmann, 2003). This interaction is mediated by residues 455H and 458E, both of which were associated with highly deleterious phenotypes, highlighting that the extended −10 motif interaction is a critically important function of RpoD in *E. coli*. In domain 4, which recognizes and binds to the −35 motif, the fitness landscape confirmed the previously known functional importance of arginine rich regions 583–589, all of which displayed highly deleterious fitness costs (Figure S7C). Additionally, RpoD also interacts with the core RNAP through various residues across all three domains, including 384L, 387V, and most residues between 402 and 413 in domain 2; 504P and 506S in domain 3; and 563F, 565I, and 598L in domain 4. Almost all of these residues, which included many branched chain amino acids, had highly or variably deleterious phenotypes, illustrating that these RNAP interaction residues are functionally important at a single-residue level. Overall, of the 63 residues compiled from the literature, we confirmed fitness defects for all residues except for 7; of these 7 residues, 1 was related to core binding (487M) (Campbell et al., 2002), 2 were related to DNA duplex binding (401F and 446Q) (Fenton et al., 2000; Zhang et al., 2012), and 4 were related to −3 and −4 non-template strand binding residues (514D, 516D, 517S, and 522F) (Zhang et al., 2012). While these residues still may be associated with proper RpoD function, we did not observe any fitness effects associated with disruption of these positions through single-residue mutagenesis. Together, these data highlight that specific residues the DNA-binding structural elements as well as core RNAP binding are major determinants of sigma factor impact on cellular fitness (Figures 2D and S7D).

Importantly, the fitness landscape also identified residues with fitness effects that were not compiled in our literature search. Domain 2 residues 408G, 411G, 415A, 431A, 450I, and 453P and domain 4 residues 576V, 577G, 582V, 590I, and 591E all exhibited highly deleterious phenotypes. For domain 2 residues, we suggest potential associated functions based on nearby annotated residues, including 408G, 411G, and 415A for core RNAP-binding-related functions; 431A for proper −10 motif DNA melting; and 450I and 453P for proper extended −10 promoter motif interaction. Domain 4 residues were all located within the helix-turn-helix motif mediating the −35 motif recognition. Furthermore, of the 76 variably deleterious residue positions identified through MAGE-seq, only 28 residue positions had been known to be functional, meaning 48 novel functional residue positions were potentially discovered. Together, our RpoD fitness landscape map paints a rich residue-level picture of protein function and evolutionary diversity.

Predicting fitness landscapes of σ^{70} orthologs

Given that the primary σ^{70} factor has rich sequence diversity across bacterial genomes yet has highly conserved essential functions, we explored ways to functionally map diverse σ^{70} orthologs at the sequence level. We assumed that within a genome, the native σ^{70} protein has evolved to a near-optimal sequence. Thus, we wondered whether the protein fitness landscapes of different σ^{70} orthologs (as measured by global cellular fitness) were similar and if we could use the *E. coli* RpoD saturation mutagenesis data to assess the fitness impact of sequence variations in different orthologs. We first took the subset of 2,833 unique σ^{70} orthologs from the ~4,700 sequence set and identified residue differences compared to the *E. coli* RpoD sequence. A total of 236,101 residue differences (i.e., variants) were found, with an average of ~83 residue variants per ortholog. We then mapped each ortholog residue variant to the *E. coli* RpoD fitness values generated from the saturation mutagenesis data to assess fitness of observed natural residue variants. In general, the majority of variants exhibited a fitness values of ≥0.95 across domains 2–4 (Figure 3A), suggesting that mostly residues with near-neutral fitness impact were observed in σ^{70} orthologs. Interestingly, ~21% of residue variants had lower fitness values (<0.95) compared to 53% of all possible variants. At the individual-domain level, domain 4 (33%) had more residues with lower fitness than domain 2 (30.4%) and domain 3 (14%) compared to all possible domain variants (63%, 78%, and 29%, respectively) (Figure S8A). This result implies that domain 4 acquired seemingly deleterious mutations more frequently than expected, as compared to domains 2 and 3. Furthermore, we find that evolutionarily distant orthologs (measured in 16S divergence) tended to accumulate more deleterious mutations (Figure 3B).

To assess how observed residue variants behave together in an ortholog, we explored a simple “additive” fitness model. We calculated the expected aggregate fitness of each ortholog by integrating the fitness values of individual residue variants together, under the assumptions that residues do not interact and that each residue’s function is independent of one another. In this simple model, the expected aggregate fitness is the product of fitness values of all observed residues in each ortholog (Figure S8B). For example, the expected aggregate fitness of an ortholog with four observed residue variants, each with a fitness value of 0.99, 1.01, 0.90, and 0.80, would be 0.72 (0.99 × 1.01 × 0.90 × 0.80). We calculated the expected aggregate fitness for the 2,833 σ^{70} orthologs and plotted these values against their sequence diversity as measured by the number of residue differences with *E. coli* RpoD (Figure 3C). Interestingly, orthologs with fewer than 20 residue differences had near-neutral expected aggregate fitness values (≥0.95). Orthologs with more diversity (>20 residue differences) had lower expected aggregate fitness values, ranging down to 0.01 for the most distant orthologs (>100 residue differences or less than ~57% sequence identity to *E. coli*). These results suggest that neutral mutations are more often observed in phylogenetically similar organisms, while more distant organisms can contain seemingly deleterious mutations. To contextualize this pattern to a naive model, we generated synthetic orthologs of varying degrees of sequence diversity to *E. coli* RpoD with random residue differences and calculated their aggregate fitness (Figure 3C). As expected, the aggregate fitness of natural orthologs was much higher than that of synthetic orthologs across all sequence diversity distances.

We also calculated domain-level aggregate fitness and found distinct domain-specific patterns. Particularly, the aggregate fitness for domain 4 reached a lower value relatively faster than domain 2, suggesting that more seemingly deleterious mutations are acquired faster (Figure S8B). This result was somewhat surprising given that overall, domain 2 mutations in general had lower individual fitness values than domain 4 mutations. Upon plotting the domain-level aggregate fitness against sequence diversity across the entire ortholog (i.e., domains 2–4), we found that domain 2, and to some extent domain 4, contained residue variants that impacted aggregate fitness in more distant orthologs (Figure 3D). Even though many orthologs had low calculated aggregate fitness, we expected all σ^{70} orthologs to be optimally fit for their respective natural genetic backgrounds (e.g., fitness of 1). Therefore, we can consider a "fitness deficit" metric for each ortholog compared to *E. coli* RpoD through a simple transformation of (1 – expected aggregate fitness). This fitness deficit can arise from differences in the fitness landscapes between the ortholog and the *E. coli* RpoD due to compensatory or epistatic mutations. We speculate that neutral mutations may alter the fitness landscape to facilitate otherwise detrimental mutations that incur prohibitive fitness costs. In turn these results suggest that domain 3 residue variants, which are observed more frequently in closely related organisms to *E. coli*, may explain the fixation of seemingly detrimental mutations observed in more distant orthologs.

Functional characterization of σ^{70} orthologs in *E. coli*

In previous studies, σ^{70} orthologs heterologously expressed on a plasmid led to changes in host gene expression, showing that orthologs could interact with non-native transcriptional machinery (Gaida et al., 2015; Tomko and Dunlop, 2017). To more rigorously measure the degree of functional conservation of σ^{70} orthologs in a non-native host, we replaced the endogenous *E. coli* RpoD with different RpoD sequences from other bacteria and measured growth and transcriptional changes in the resulting strains. For σ^{70} orthologs with high functional conservation to *E. coli* σ^{70}, we expected minimal growth and transcriptional changes. On the other hand, differences in growth rate and transcriptional responses would reflect differences in σ^{70} function. We used recombineering and CRISPR selection to replace the chromosomal *E. coli* RpoD with orthologs from diverse bacteria (see STAR Methods). *E. coli* mutants (EcJP1–15), each carrying one of 15 orthologs with 3- to 96-residue differences from *E. coli* RpoD, were successfully generated (Table 1). These orthologs represented a diverse panel of σ^{70} sequences mostly belonging to Proteobacteria (six Gammaproteobacteria, four Alphaprotepbacteria, three Betaproteobacteria, and one Deltaproteobacteria) and one Actinobacteria (Figure S9A) and contained residue differences of varying fitness impacts (Figure 4A).

To measure the global fitness impact of each ortholog on *E. coli* growth, we pooled all strains with the WT *E. coli* and grew them in a turbidostat. Sampling from the population over time and amplicon sequencing the *rpoD* region yielded relative fitness measurements of each σ^{70} ortholog, which showed high correlation across two independent competition assays, and also matched growth rates derived from individual growth assays (Figure S9B). We compared the measured fitness with the expected aggregate fitness derived from the saturation mutagenesis data and found a good positive correlation (Figures 4B and S9C). We noted that the measured fitness values were generally higher than the expected aggregate

fitness values, which implied that residue differences in orthologs had positive synergistic effects on fitness. Importantly, this fitness differential (i.e., fitness difference between expected aggregate fitness and measured fitness) positively correlated with the number of residue differences (Figure 4C). Furthermore, we analyzed the evolutionary coupling scores for all pairs of residue positions in *rpoD* in an attempt to identify strongly coupled residue positions that might explain the differences between measured and predicted fitness. While we observed that the sum of evolutionary coupling scores of orthologs correlated with fitness differentials as expected, no individual residue pairs made significant contributions to the coupling score sum, suggesting that it was the cumulative effects of weakly coupled residue pairs that yielded the observed fitness differentials (Figure S9D). These results demonstrate that distant orthologs can function in a non-native host and that synergistically beneficial interactions between residues likely buffer against otherwise deleterious mutations as sequences diverge over time.

To further probe the effects of σ^{70} orthologs on cellular fitness, we performed detailed gene expression profiling on two strains (EcJP9 and EcJP14) that contained a σ^{70} ortholog derived from either *Myxococcus xanthus* (Mx σ^{70}) or *Oligella urethralis* (Ou σ^{70}). While these orthologs had a similar number of residue differences from *E. coli* σ^{70} (51 for Mx σ^{70} and 50 for Ou σ^{70}), their measured fitness differentials were quite different. Mx σ^{70} exhibited a low measured fitness (~0.49), similar to its expected aggregate fitness (~0.48). In contrast, Ou σ^{70} exhibited a high measured fitness (~0.95), while its expected aggregate fitness was much lower (~0.71). We therefore performed RNA sequencing (RNA-seq) of the Mx σ^{70} and Ou σ^{70} strains to profile their transcriptional changes compared to WT *E. coli* (Ec σ^{70}) across biological replicates. Interestingly, compared to Ec σ^{70}, the number of differentially expressed genes (DEGs) for Ou σ^{70} was fewer than the number of DEGs for Mx σ^{70} (Figures 4D and S10A), in line with what might be expected from the measured fitness data. Accordingly, the transcriptome of Ou σ^{70} was also more similar to Ec σ^{70} than that of Mx σ^{70} (Figure S10B).

In general, about three times more DEGs were observed in Mx σ^{70} than in Ou σ^{70}, but there was not a big difference in the direction of gene expression change (approximately half were upregulated and half were downregulated) (Figures 4D and S10C). Among the ~300 essential *E. coli* genes (Baba et al., 2006), both Mx σ^{70} and Ou σ^{70} upregulated a similar number of essential genes (8 for Mx and 5 for Ou), while Mx σ^{70} downregulated 33 essential genes compared to just 2 in Ou σ^{70} (Figure 4E). The larger number of downregulated essential genes in Mx σ^{70} likely contributed to the significant fitness decrease observed in the strain. While it might have been expected that the transcriptional differences in Mx σ^{70} and Ou σ^{70} could be due to differences in their respective optimal σ^{70} binding motif, we could not identify any significant motif differences from the regulatory regions of their DEGs. Together, these results highlight the complex functional and fitness constraints that shape σ^{70} evolution in bacteria.

Transcriptional activation potential of σ^{70} orthologs

To more deeply profile the global transcriptional changes due to different σ^{70} factors in *E. coli*, we utilized a multiplexed reporter assay to characterize the activity of a library of

diverse regulatory sequences in strains EcJP9 (Ou σ^{70}), EcJP14 (Mx σ^{70}), and EcNR2 (Ec σ^{70}). Measurement of non-native regulatory sequences in *E. coli* that possess alternative σ^{70} orthologs allowed us to directly determine differences in σ^{70} specificity or function independent of the endogenous regulatory network. We thus mined for 5′ intergenic sequences that are >100 bp in the *E. coli*, *M. xanthus*, and *O. urethralis* genomes to yield a library of ~6,000 regulatory sequences (Figure S11A; see STAR Methods). This library was synthesized as a pool of barcoded oligonucleotides, cloned into a reporter vector, and transformed into strains EcJP9, EcJP14, and EcNR2, which possessed different σ^{70} factors (Figure S11B). We next performed targeted DNA sequencing (DNA-seq) and RNA-seq to yield relative abundance measurements of DNA and RNA levels that were then used to compute a relative transcription activity (Tx value) for each regulatory sequence. Tx values were normalized to qPCR-derived expression levels of an invariant control gene, *infC*, which enabled comparisons of Tx activity between different *E. coli* strains (i.e., Tx_O, Tx_M, and Tx_E). These Tx measurements showed high correlation between replicates (both biological and barcode replicates) (Figure S11C) and with qPCR-based expression measurements (Figure S11D) and spanned over three orders of magnitude.

Comparing σ^{70} orthologs, we observed that the median Tx_M was lower than the median Tx_E or median Tx_O (Figure 5A), suggesting that Mx σ^{70} had lower transcriptional output overall, which is in agreement with its observed reduced fitness (~0.49). Interestingly in all orthologs, regulatory sequences derived from the *Ou* genome had the highest expression, followed by *Ec*- and *Mx*-derived sequences. Given the genomic GC content of *Ou*, *Ec*, and *Mx* (46%, 50%, and 69%, respectively), our observed Tx trends are in agreement with previous studies that found that lower GC regulatory sequences yield higher gene expression (Chen et al., 2007; Johns et al., 2018; Figure S12B). To understand how the σ^{70} orthologs' Tx activity changed compared to WT *E. coli*, we normalized Tx_O values for each regulatory sequence with their corresponding Tx_E values ($Tx_{O/E}$) and similarly for Tx_M with Tx_E ($Tx_{M/E}$) and further sub-grouped these values by the genomic source of the regulatory sequences (i.e., $Tx_{O/E}{}^O$, $Tx_{O/E}{}^E$, $Tx_{O/E}{}^M$, etc.). We observed that $Tx_{O/E}{}^M$ was notably higher than $Tx_{O/E}{}^O$, and $Tx_{O/E}{}^E$ (Kolmogorov-Smirnov [KS] test, $p < 10^{-104}$ and $p < 10^{-168}$, respectively) (Figure 5B), indicating that *Mx* sequences had a higher activity with *Ou* σ^{70} than *Ec* σ^{70}. We also observed small but statistically significant differences in the distributions of $Tx_{M/E}{}^O$, $Tx_{M/E}{}^E$, $Tx_{M/E}{}^M$ (KS test: $Tx_{M/E}{}^M$ and $Tx_{M/E}{}^E$, $p < 10^{-22}$; $Tx_{M/E}{}^M$ and $Tx_{M/E}{}^O$, $p < 10^{-22}$; $Tx_{M/E}{}^O$ and $Tx_{M/E}{}^E$, $p < 0.0164$), indicating not only that GC content inversely affected expression within each σ^{70} orthologs as previously known, but also that the magnitude of GC content effect seemed to vary by σ^{70} ortholog.

To better understand the differences in the relationships between GC content of regulatory sequences and the resulting gene expression patterns from different σ^{70}s, we performed a linear regression to identify correlations between GC content and Tx values between σ^{70} orthologs. The slope of the Mx σ^{70} regression indicated the strongest inverse relationship between GC content and Tx levels, while the Ou σ^{70} slope showed the weakest relationship (t test, $p < 0.05$ for all pairwise slope comparisons) (Figure 5C). This result suggested that the same degree of decrease in regulatory sequence GC content generally yielded a larger transcriptional increase in Mx σ^{70} compared to Ec σ^{70} or Ou σ^{70}. Interestingly, this relationship also correlated with genomic GC content, implying that the gene expression

patterns from higher GC genomes may be more sensitive to GC content in the regulatory region. We speculated that this GC-dependent sensitivity may be due to the fact that the canonical σ^{70} motif is rich in A/T bases. In a GC-rich genome, simply having higher AT content in intergenic regions may be sufficient to facilitate recognition and binding. In contrast, AT-rich genomes may require a more stringent sequence similarity to the optimal σ^{70} motif for recognition and binding by σ^{70} factors, as there is an abundance of A/T bases throughout. Therefore, in GC-rich genomes, σ^{70} binding sites should have more A/T bases than the background GC distribution of intergenic regions and a higher variance in GC content in the intergenic regions than in AT-rich genomes. Indeed, when we mined 5′ upstream intergenic regions of coding sequences from ~1,300 bacterial genomes (see STAR Methods) and compared the GC content against the GC variance, we observed significantly higher variance in GC content for GC-rich genomes (Figure S12C). Together, these results suggest that higher GC variance in intergenic regions may be a consequence of σ^{70} evolution that is influenced by genomic GC content of the organism.

DISCUSSION

Here, we explored the evolutionary sequence diversity of the primary σ^{70} and its fitness landscape in *E. coli* and compared its functional capacity with σ^{70} orthologs. While there is generally high evolutionary conservation of σ^{70}, domain-level differences suggested different evolutionary forces driving diversification of this global regulator. Using MAGE-seq, we generated a saturation mutagenesis library that tiled across domains 2–4 of the *E. coli* σ^{70} and characterized the resulting fitness landscape. By contextualizing evolutionary sequence divergence with the *E. coli* σ^{70} fitness landscape, we found that *E. coli* σ^{70} tolerated most individual residue differences found in natural orthologs. Interestingly, residues that incurred significant fitness costs were observed in orthologs that were phylogenetically distant from *E. coli*, suggesting reshaped fitness landscapes that compensated for these otherwise predicted fitness deficits. Accordingly, when natural orthologs were used in place of the endogenous *E. coli* σ^{70}, we found that fitness losses were generally lower than predicted from a simple aggregate fitness model. Decreased cellular fitness with different σ^{70} orthologs could be attributed to downregulation of essential genes based on transcriptomic measurements. Finally, we used a regulatory sequence library to identify differences in regulatory activation capacity of two σ^{70} orthologs compared to *E. coli* σ^{70} and identified unique patterns of expression that were dependent on both regulatory GC content and the source species of the σ^{70} orthologs.

Bacterial genomic GC content appears to correlate with genome size and regulatory complexity (McCutcheon and Moran, 2011). Genomic GC drift is thought to arise from mutational processes and selective biases (Hershberg and Petrov, 2010; Hildebrand et al., 2010; Musto et al., 2006; Raghavan et al., 2012) that also affect GC content of intragenic UTRs (Muto and Osawa, 1987). In this study, we observed that σ^{70} orthologs derived from *Myxococcus* and *Oligella* species with very different GC content exhibited distinct patterns of transcription activation potential. While σ^{70} appeared to be largely functionally conserved and portable between bacteria within the phylum level, σ^{70} of GC-rich *Myxococcus* yielded a stronger inverse relationship between regulatory GC content and expression levels than that of σ^{70} of AT-rich *Oligella*. GC-rich genomes were also observed to have UTRs with

larger variances in GC content, which further supports the link between the magnitude of GC content effect on transcription and σ^{70} functionality.

In this study, fitness effects of single amino acid mutations were used in a simple additive model, assuming each mutation contributes to fitness independently, to extrapolate fitness of orthologs with multiple amino acid differences. In reality, epistasis, the notion that functional consequences (e.g., fitness) of amino acid changes depend on the specific protein sequence context, is likely involved in σ^{70} evolution (Starr and Thornton, 2016; Storz, 2018). The role of epistasis is suggested in the observation that in orthologs with few amino acid differences to *E. coli* σ^{70} (more similar protein sequence context), fitness effects of amino acid differences were neutral in the *E. coli* σ^{70} sequence background. Conversely, in orthologs with many amino acid differences (less similar protein sequence context), fitness effects of amino acid differences were more deleterious in the *E. coli* σ^{70} sequence background. Furthermore, fitness differentials between predicted and measured fitness of the 15 ortholog mutant strains are also suggestive of possible epistatic effects. Efforts to more systematically profile epistasis, such as using pairwise mutagenesis or saturation mutagenesis across multiple sequence backgrounds, will yield more insights into the relationship between epistasis and σ^{70} evolution. Lastly, σ^{70} is a DNA-binding protein, unlike many enzymes that have substrates that do not evolve. Therefore, a complete model of σ^{70} fitness would not only incorporate epistasis but also accurately incorporate covariation and co-evolution of cognate promoter sequences.

For practical reasons, we mainly focused on domains 2–4 of σ^{70} in our study. However, domain 1, which is more variable than domains 2–4, may further contribute to shaping the evolutionary trajectory of σ^{70} not accounted for here. Another caveat of this study is that the σ^{70} orthologs were characterized in the context of *E. coli* and that suboptimal interactions of specific residues in σ^{70} orthologs with the native *E. coli* RNAP could impede the global transcriptome in very complex ways. For instance, Mx σ^{70} exhibited a lower overall transcription level based on our promoter library measurements, which may have been caused by one of its many residue differences from *E. coli* σ^{70}, potentially reducing its ability to bind or interact with the *E. coli* RNAP. Exploration of other bacterial backgrounds could shed light on host-specific differences to better explore σ^{70} orthologs beyond Proteobacteria. In this study, efforts to introduce σ^{70} variants from more phylogenetically distant bacteria were mostly unsuccessful, highlighted by significant fitness costs and functional differences that may be found in these more distant σ^{70} orthologs. Lastly, we note that MAGE-seq protocol used here only allowed grow at 30°C (i.e., the recombineering system is induced at 37°C–42°C). As environmental selection pressures can dictate fitness, some degree of fitness variations may be observed at different growth temperatures; MAGE-seq using arabinose-inducible pKD46 (Datsenko and Wanner, 2000) could enable fitness measurements at other growth temperatures. Further explorations could propel understanding and better modeling of bacterial regulation to allow precise control and engineering of gene regulation in a variety of non-model bacteria while accounting for complex evolutionary forces driving the selection of global regulators.

STAR⋆METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Harris Wang (hw2429@columbia.edu).

Materials availability—Requests for plasmids and strains described in this study can be made to the Lead Contact, Harris Wang (hw2429@columbia.edu).

Data and code availability

- All sequencing data have been deposited at ArrayExpress and are publicly available as of the date of publication. Accession numbers are listed in the Key resources table.
- Code used for analysis are publicly available as of the date of publication. Github repository links are listed in the Key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bacteria culture—*E. coli* MegaX DH10B strain was cultured in liquid medium (LB) at 37°C. *E. coli* EcNR2 strain and all its derivatives were cultured at 30°C.

METHOD DETAILS

Saturation Mutagenesis of *E. coli* RpoD—*E. coli* EcNR2 strain was used to generate a saturation mutagenesis library of *rpoD* using MAGE. 236 70-mer oligonucleotides were designed to systematically mutate all positions between 379–613 (which covers Domains 2–4) and the stop codon of *E. coli rpoD* gene (Synthego). Each oligonucleotide was designed with 34 and 33 base pairs of 5′ and 3′ homology to *rpoD* respectively. Homology regions flanked an NNN codon for mutagenesis of all 64 codon variants. Mutagenesis region (positions 379–613) was split into six bins and MAGE oligos were pooled accordingly to achieve full sequencing coverage. First two bins covered 44 and 48 amino acids, respectively, while the subsequent four bins covered 36 amino acids each. Six iterative rounds of MAGE were carried out. EcNR2 strain was inoculated from a glycerol stock and cultured overnight at 30°C. Next morning, 100uL of the overnight culture was inoculated into 3mL of LB with 50 ug/mL carbenicillin (Fisher Scientific, BP26485) and grown until 0.5 OD_{600} was reached. Then the cultures were transferred into a 42°C shaking water bath for 15 minutes to induce recombineering proteins. Following induction, the culture was chilled down immediately in an ice-slurry. 1mL of cells were pelleted in a pre-chilled centrifuge (Eppendorf, 5424R) and then washed twice with pre-chilled distilled water (ThermoFisher, 15230162). Washed cell pellet was resuspended in 50uL of 5uM MAGE oligo pool, transferred to a 0.1cm electroporation cuvette (Bio-Rad, 1652089), electroporated at 1.8kv (Bio-Rad, Micro-Pulser), and recovered in 3mL LB at 30C until 0.5 OD_{600} was reached. This protocol was repeated until 6 total MAGE cycles were

performed at which point resulting populations were transferred immediately to a turbidostat for competition experiments.

Competition Experiments—Competition experiment with the saturation mutagenesis library was performed on a custom built turbidostat in a 30°C incubator. A single competition was performed for each MAGE mutagenesis bin. An LED and a photodiode were used to monitor the optical density of the culture over time. When an OD_{600} threshold of 0.4 was reached, turbidostat automatically diluted the culture by half to OD_{600} of 0.2 through peristaltic pumps that added fresh LB and removed excess media. This ensured that the competition cultures were constantly maintained at exponential growth phase, between OD_{600} 0.2–0.4. Cells from the final MAGE round were inoculated into a culture tube containing 10mL LB and then placed in the turbidostat. First time point was collected when the population reached 0.4 OD_{600} for the first time. Subsequent time points were collected at 1, 2, 3, 4, 5, 6, 9, and 12 hours after the initial time point for a total of 9 time points. To collect time point samples, we removed 1 mL of the culture with using a 1mL Luer-Lok syringe (BD, 309628) with a blunt needle (Air-Tite, NB18212), pelleted in a pre-chilled centrifuge (Eppendorf, 5424R), washed once with pre-chilled PBS (GIBCO, 10010049), pelleted again, removed supernatant, and stored the pellet at −20°C until all samples were collected. For competition experiments with *E. coli* strains expressing ortholog σ^{70} variants, we used the eVOLVER (Wong et al., 2018) on the turbidostat mode. All parameters for the turbidostat were kept same. Overnight cultures of sigma factor ortholog variants were pooled to equal volume and used to inoculate an eVOLVER culture tube containing 20mL LB. Samples were harvested using the same regimen as described above for MAGE mutants.

Library Preparation and Sequencing—Genomic DNA was prepped from each time point sample (GE life sciences, 28904259). 1uL gDNA (~20ng) was used to amplify the mutagenesis loci of each MAGE bin in a 20uL PCR reaction with 1x Q5 Hot start HiFi Master Mix (NEB, M0543L), 1x SYBR Green (Invitrogen, S7567) and 0.5uM of forward and reverse primer pools (Data S2). PCR (95°C 30 s, cycle: 95°C 10 s, 65°C 10 s, 72°C 10 s; and 72°C 2min) was performed on a real time PCR machine (Bio-Rad, CFX-96) and the reaction was terminated during exponential amplification. Same PCR steps were used to amplify gDNA prepped from ortholog competition samples using primers (Data S2) designed to amplify a region (corresponding to RpoD residues 532–581) that could differentiate all ortholog sequence variants. 0.1uL of the first PCR was used to perform a second 20uL (1x Q5 Hot start HiFi Master Mix, 1x SYBR green, 0.5uM of p5_X and p7_X amp2 primers;Data S2) PCR (95°C 30 s, cycle: 95°C 10 s, 72°C 30 s; and 72°C 2min) reaction (ran on real time PCR machine to terminate reaction during exponential amplification) to add sample barcode indexes and Illumina p5 and p7 adaptor sequences. Samples were pooled together for sequencing following quantification of dsDNA concentration (Invitrogen, Q32851) of each sample, cleaned up using 2x SPRI beads (Beckman Coulter, A63881), and sequenced according to Illumina sequencing protocols. Three Illumina NextSeq 300-cycle (150 pair-end) mid-output kits were used to sequence six RpoD MAGE mutagenesis bins (two bins each) (Illumina, 20024905). Ortholog competition library was sequenced using an Illumina MiSeq 300-cycle micro-output mode (200 single-end) (Illumina, MS-103–1002).

Cloning RpoD Ortholog Sequence Strains—To generate orthologous sigma factor strains we first generated an EcNR2 strain lacking carbenicillin resistance by inserting three stop codons and a frameshift mutation into the *bla* gene through MAGE (EcJP0) (Wang and Church, 2011). Next, plasmid pMA7CR encoding an inducible cas9 gene was introduced into EcNR2 through electroporation using standard transformation protocols. Plasmids encoding gRNA targeting different *rpoD* loci (rpoD_pam0, rpoD_pam1, rpoD_pam4, rpoD_pam6) were cloned into pMAZ-SK plasmid using USER cloning as previously described (Ronda et al., 2016) (pMAZ-rpoD_pam0, pMAZ-rpoD_pam1, etc.). Dual gRNA plasmid constructs were cloned through Gibson assembly (NEB, E5520S) of gRNA expression cassette of one gRNA plasmid into a linearized plasmid encoding a different gRNA sequence. Specifically, pMAZ-rpoD_pam14 plasmid was made through Gibson assembly of pam1 expression loci (amplified from pMAZ-rpoD_pam1 plasmid with primers JP559 and JP560 and pam4 linearized plasmid (amplified from pMAZ-rpoD_pam4 plasmid with primers JP561 and JP562). pMAZ-rpoD_pam06 gRNA plasmid was constructed using the same approach with pMAZ-rpoD_pam0 and pMAZ-rpoD_pam6 plasmids in place of pMAZ-rpoD_pam1 and pMAZ-rpoD_pam4 plasmids respectively.

Ortholog sigma factor sequence variants were synthesized as dsDNA fragments (IDT, gBlocks). To generate orthologous sigma factor strains, dsDNA fragment encoding sigma factor variant and a dual gRNA plasmid (pMAZ-rpoD_pam14 for all fragments except for construct F9Y183 which was cloned with pMAZ-rpoD_pam06) were electroporated together into EcNR2 strain with pMA7CR. Following electroporation, samples were inoculated into 3mL LB + 100ug/mL carbenicillin and recovered at 30°C for 1 hour. Then, kanamycin (Fisher Scientific, BP9065) was added to the culture at 50ug/mL final concentration and was recovered for another 2 hours at 30°C. Then, anhydrous tetracycline (Cayman, 10009542) was added to the culture at 200ng/mL final concentration and recovered for another 2 hours at 30°C. Dilutions of the recovered culture was used to plate on LB-agar plates with 100ug/mL carbenicillin and 50ug/mL kanamycin and was incubated at 30°C overnight. Recombinant RpoD clones were screened via sanger sequencing with primers JP130 and JP131.

Regulatory Sequence Library Construction—Intergenic regions from 5′ upstream of start codons of every annotated gene coding sequence were mined from the genomes of *Escherichia coli* (NC_000913), *Myxococcus xanthus* (NC_008095), *Oligella urethalis* (NZ_AQVB00000000.1). Upstream sequences shorter than 100 base pairs were discarded and 100 base pairs directly upstream of each start codon of the remaining sequences were compiled. To each regulatory sequence, we added a start codon, a unique 12-mer barcode (> 1 hamming distance to all other barcodes), and flanking restriction digest cut sites (BamHI and PstI) and common amplification sequences to yield a final 165bp construct. Each regulatory sequence was synthesized twice with two unique barcodes to yield a 12,254-member library which was synthesized as an oligonucleotide pool (Agilent, G7721A). The oligo pool was amplified in 16 parallel 20uL reactions (1x Q5 Hot start HiFi Master Mix (NEB, M0543L), 0.5uM each primer JP194, JP195) for 7 cycles to prevent overamplification (95°C 30 s, 7 cycles: 95°C 10 s, 72°C 30 s; and 72°C 2min). PCR reactions were pooled and cleaned up with beads (Beckman Coulter, A63881). Purified

library and pNJ7 plasmid (Johns et al., 2018) were digested with BamHI (NEB, R0136M) and PstI (NEB, R0140M), PCR purified (Zymo, D4033), and ligated with T4 DNA ligase (NEB, M0202M). Resulting ligation reaction was PCR purified, mixed into a 100uL aliquot of *E. coli* MegaX DH10B electrocompetent cells (Invitrogen, C640003), and aliquoted into four prechilled 1mm cuvettes (BioRad, 1652089) and electroporated. Following recovery, we used 5uL of the culture and plated dilutions to quantify cloning coverage (> 1000x cloning coverage). Rest of the recovery culture was inoculated to 1:25 ratio into LB with 20ug/mL chloramphenicol (Sigma, C0378). Following an overnight incubation, 1mL of the culture was used to inoculate 100mL LB with 20ug/mL chloramphenicol. The culture was incubated until mid-log growth was reached (OD600 0.5) and 50mL was used to prep plasmid DNA (Zymo, D4200). Rest of the culture was used to generate frozen stocks. Plasmid DNA was used to transform EcJP9, EcJP14 and EcNR2 strains at > 100x coverage.

Regulatory Sequence Library Sequencing Preparation—Overnight culture of each library was prepared by adding 1mL of thawed frozen library glycerol stock to 25mL LB and was grown overnight at 30C. 2mL of the overnight culture was used to inoculate 60mL LB and was grown until mid-log phase was reached. 5mL of the culture was used to isolate plasmids (QIAGEN, 27106). Rest of the culture was pelleted, washed once with PBS and harvested for RNA using the RNAsnap (Stead et al., 2012) protocol and cleaned up with RNA clean and concentrator kit (Zymo, R1018). DNA library was prepared through the same two-step amplification as MAGE libraries using 1uL of plasmid miniprep. RNA library was prepared by first digesting DNA with turbo DNase and cleaned up using RNA clean and concentrator. RNA samples were reverse transcribed with Maxima H minus reverse transcriptase (Thermo Scientific, EO0751) using gene specific primers against *sfGFP* reporter gene (12.5uL of RNA, 1uL of primer JP750, 1uL of 10mM dNTPs incubated at 65C for 5 minutes, then on ice 1min, then add 4uL 5x RT buffer, 0.5uL RNase inhibitor (Thermo Scientific, EO0381), 1uL Maxima RT, then incubate with the following protocol: 42°C 90 minutes, cycle 9 times: 50°C 2 minutes, 42°C 2 minutes; 85°C 5 minutes, 4°C hold). 1uL RNase A (Thermo Scientific, EN0531) and 1uL RNase H (NEB, M0297S) were added to the reaction and incubated for 30 minutes at 37°C. Then, bead cleanup was used to purify cDNA. Adaptor was ligated using T4 RNA ligase (NEB, M0437M) (5.1uL cDNA, 2uL 40mM DNA adaptor oligo, incubate 75°C for 3 minutes, 1 minute on ice, add 2uL 10x T4 RNA ligase buffer, 0.8uL DMSO, 0.2uL 100mM ATP, 8.4uL 50% PEG, 1.5uL T4 RNA ligase, and incubate at 22C for 16 hours). Adaptor ligated cDNA samples were then purified with beads. Sequencing library was then prepared through a two-step amplification, using the same protocol as for DNA samples. For measuring expression of regulatory element isolates, total RNA was harvested from a 5mL cell culture in mid log growth phase. cDNA was prepped using the same protocols as above. qPCR was performed with primers against *sfGFP* reporter gene and *infC* gene was used as a reference house-keeping gene.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evolutionary Sequence Divergence Analysis—Orthologs of *E. coli* RpoD were initially mined using the KEGG (Kanehisa et al., 2016) orthology database (K03086, RNA polymerase primary sigma factor). We refined the ortholog set by limiting the scope to bacterial proteins that also encoded the following Pfam (El-Gebali et al., 2019) domains:

Sigma-70 region 2 (PF04542), Sigma-70 region 3 (PF04539), Sigma-70 region 4 (PF04545). Furthermore, alternative sigma factors entries (e.g., rpoH, sigS, rpoS) were removed from the ortholog set. Sequences and corresponding UniProt accessions were downloaded on 2020–08-25 to yield a final ortholog set of 4,703 sequence variants. Taxonomic data of each orthologs were extracted from each UniProt entry's metadata. GC contents of each coding sequences were extracted from parsing through linked nucleotide refseq accessions. Amino acid sequences of orthologs were used to generate a multiple sequence alignment (MSA) with Clustal Omega (Sievers et al., 2011) using the following parameters:–full,–full-iter,–iter = 5. Resulting MSA was used to quantify σ^{70} conservation via Jensen-Shannon divergence (Capra and Singh, 2007). Maximum Likelihood phylogenetic tree was constructed using FastTree2 (Price et al., 2010). Phylogenetic tree was visualized through iTOL (Letunic and Bork, 2019). We extracted RpoD evolutionary divergence distances using the branch length distances between *E. coli* RpoD and all other orthologs on the phylogenetic tree. Corresponding 16S sequence for each UniProt RpoD ortholog entry was extracted from Greengenes database by matching the NCBI taxa IDs from Greengenes accessions and Uniprot metadata for each entry. Compiled 16S sequences were aligned with MAFFT (Katoh et al., 2002) using default settings. Resulting MSA was used to generate a phylogenetic tree with FastTree2. 16S evolutionary divergence distances were compiled using the branch length distance between *E. coli* 16S and all other 16S sequences. To normalize for uneven phylogenetic sampling bias in the databases and account for over-represented sequences, we collapsed sequences down to unique Domain 2–4 sequences to yield a set of 2,833 sequence variants which were used to study domain level sequence divergence.

σ^{70} Sequence Motif Analysis—To screen for genes regulated by primary σ^{70} we sought to identify cellular functions under σ^{70} regulation. From regulonDB (Gama-Castro et al., 2016), we isolated all genes regulated by *E. coli* σ^{70} and compared their COG functional categories (Galperin et al., 2015) against all *E. coli* genes and found that F/K COGs (nucleotide metabolism and transcription categories, respectively) were enriched in σ^{70} regulated genes (Figure S3E). Next, we mined 5′ upstream regions (from 25 to 100 base pairs upstream of start codons) of all F/K COG genes in each bacterial genome in the COG dataset. BioProspector (Liu et al., 2001) was used to generate bipartite sequence motifs for each genome using the following standard parameters: -d 1 -n 200 -w 8 -W 8. Each motif search was background normalized using its cognate genomic sequence. 12 motifs were generated for each genome with varying maximum and minimum gap parameters (min gap: 13–16, max gap: 19–21) and the best motif from each set was selected by comparing the sum of all motif correlations against all other motifs. Next, each motif was subjected to score (BioProspector score > 2) and counts (upstream region counts > 50) thresholds. Lastly, we selected a subset of motifs from genomes which primary σ^{70} were present in our σ^{70} ortholog dataset, yielding a final set of 188 motif and σ^{70} ortholog pairs.

Saturation Mutagenesis Library Sequencing Analysis—Raw reads from each sample were pair-end merged with SeqPrep using default settings. Next, expected error score was calculated for each read and any reads with expected error score > 1 was designated as low quality read and then discarded from further analysis. Then each high-quality read was

tallied as a WT sequence or a mutant sequence. Any sequences with mutations in more than one codon were discarded. Counts of each sequence variants were used to calculate relative frequencies of WT and mutants. To correct for miscalls, relative frequencies of mutants from sequencing of the control WT was used to subtract the relative frequencies of mutants from each time point. For each mutant sequence, corrected relative frequencies from each time points were used to generate a log linear regression. The slope of the regression was normalized to the dilution rate of the culture during the competition experiment to yield a fitness metric from 0 to 1. Fitness of 1 means that the mutant has the same growth rate as the WT sequence while fitness of 0 means that the mutant does not grow.

Transcriptomic analysis of orthologous rpoD mutants—EcNR2 strains encoding three different σ^{70} sequence variants (Ec σ^{70}, Mx σ^{70}, Ou σ^{70}) strains were grown overnight from a glycerol stock. 166uL of the overnight culture was used to inoculate a 5mL culture and was harvested for total RNA when the culture reached mid-log growth (OD600 0.5). For each strain, a total of four biological replicates were harvested across two independent days via RNAsnap (Stead et al., 2012). DNA was removed from the total RNA with turbo DNase (Invitrogen, AM2239). Next, rRNA was depleted with the Ribo-Zero magnetic kit for Bacteria (Illumina, MRZB12424). DNA free, rRNA depleted RNA samples were used to prep a sequencing library with NEBnext ultra directional RNA library kit (NEB, E7420L). Sequenced on the Illumina Nextseq platform with 300 cycle mid output kit. Analysis of resulting RNaseq data was carried out with Trimmomatic (Bolger et al., 2014) for cleaning up reads, bowtie (Langmead et al., 2009) for alignment, HTSeq (Anders et al., 2015) for RNA counts, and DEseq2 (Love et al., 2014) for differential gene expression analysis.

Regulatory Sequence Library Analysis—Raw sequencing reads were pair end merged using SeqPrep. Then using a custom python script, merged reads with low quality scores were removed (expected error > 2 for the full merged read). Next, counts of each regulatory sequence construct with correct barcode identifiers were tallied with up to 4% mismatch tolerance in regulatory sequence regions and no mismatch allowed in barcode regions. Counts of each construct were divided by the total sum of all constructs to yield relative abundance measurements. For constructs with 10+ DNA and RNA counts, we calculate a Tx value by dividing its relative RNA abundance by its relative DNA abundance. To enable comparison of Tx values between samples, Tx values were normalized using the qPCR expression ratios.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank members of the Wang lab for helpful scientific discussions and feedback. H.H.W. acknowledges funding support from the NIH (U01GM110714–01A1), NSF (MCB-1453219, MCB-2032259), Sloan Foundation (FR-2015–65795), DARPA (W911NF-15–2-0065), ONR (N00014–15-1–2704), and the Irma T. Hirschl Trust. We thank Synthego for providing oligos for MAGE-seq experiments.

REFERENCES

Alper H, and Stephanopoulos G (2007). Global transcription machinery engineering: a new approach for improving cellular phenotype. Metab. Eng 9, 258–267. [PubMed: 17292651]

Anders S, Pyl PT, and Huber W (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. [PubMed: 25260700]

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, and Mori H (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol 2, 0008. [PubMed: 16738554]

Bae B, Feklistov A, Lass-Napiorkowska A, Landick R, and Darst SA (2015). Structure of a bacterial RNA polymerase holoenzyme open promoter complex. eLife 4, e08504.

Bolger AM, Lohse M, and Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. [PubMed: 24695404]

Bottacini F, Zomer A, Milani C, Ferrario C, Lugli GA, Egan M, Ventura M, and van Sinderen D (2017). Global transcriptional landscape and promoter mapping of the gut commensal Bifidobacterium breve UCC2003. BMC Genomics 18, 991. [PubMed: 29281966]

Browning DF, and Busby SJW (2016). Local and global regulation of transcription initiation in bacteria. Nat. Rev. Microbiol 14, 638–650. [PubMed: 27498839]

Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, and Darst SA (2002). Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Mol. Cell 9, 527–539. [PubMed: 11931761]

Capra JA, and Singh M (2007). Predicting functionally important residues from sequence conservation. Bioinformatics 23, 1875–1882. [PubMed: 17519246]

Chen S, Bagdasarian M, Kaufman MG, Bates AK, and Walker ED (2007). Mutational analysis of the ompA promoter from Flavobacterium johnsoniae. J. Bacteriol 189, 5108–5118. [PubMed: 17483221]

Datsenko KA, and Wanner BL (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. USA 97, 6640–6645. [PubMed: 10829079]

de Visser JAGM, and Krug J (2014). Empirical fitness landscapes and the predictability of evolution. Nat. Rev. Genet 15, 480–490. [PubMed: 24913663]

Domínguez-Cuevas P, and Marqués S (2004). Compiling Sigma-70-Dependent Promoters. In Pseudomonas (Springer US), pp. 319–343.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. 47 (D1), D427–D432. [PubMed: 30357350]

Feklistov A, and Darst SA (2011). Structural basis for promoter-10 element recognition by the bacterial RNA polymerase σ subunit. Cell 147, 1257–1269. [PubMed: 22136875]

Feklístov A, Sharon BD, Darst SA, and Gross CA (2014). Bacterial sigma factors: a historical, structural, and genomic perspective. Annu. Rev. Microbiol 68, 357–376. [PubMed: 25002089]

Feng Y, Zhang Y, and Ebright RH (2016). Structural basis of transcription activation. Science 352, 1330–1333. [PubMed: 27284196]

Fenton MS, Lee SJ, and Gralla JD (2000). Escherichia coli promoter opening and −10 recognition: mutational analysis of sigma70. EMBO J. 19, 1130–1137. [PubMed: 10698953]

Firnberg E, Labonte JW, Gray JJ, and Ostermeier M (2014). A comprehensive, high-resolution map of a gene's fitness landscape. Mol. Biol. Evol 31, 1581–1592. [PubMed: 24567513]

Gaida SM, Sandoval NR, Nicolaou SA, Chen Y, Venkataramanan KP, and Papoutsakis ET (2015). Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. Nat. Commun 6, 7045. [PubMed: 25944046]

Galperin MY, Makarova KS, Wolf YI, and Koonin EV (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43, D261–D269. [PubMed: 25428365]

Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, et al. (2016).

RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 44 (D1), D133–D143. [PubMed: 26527724]

Gardella T, Moyle H, and Susskind MM (1989). A mutant Escherichia coli σ 70 subunit of RNA polymerase with altered promoter specificity. J. Mol. Biol 206, 579–590. [PubMed: 2661827]

Grigorova IL, Phleger NJ, Mutalik VK, and Gross CA (2006). Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. Proc. Natl. Acad. Sci. USA 103, 5332–5337. [PubMed: 16567622]

Gruber TM, and Bryant DA (1997). Molecular systematic studies of eubacteria, using sigma70-type sigma factors of group 1 and group 2. J. Bacteriol. 179, 1734–1747. [PubMed: 9045836]

Hershberg R, and Petrov DA (2010). Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6, e1001115. [PubMed: 20838599]

Hildebrand F, Meyer A, and Eyre-Walker A (2010). Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 6, e1001107. [PubMed: 20838593]

Hook-Barnard IG, and Hinton DM (2007). Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. Gene Regul. Syst. Bio. 1, 275–293.

Ishihama A (2000). Functional modulation of Escherichia coli RNA polymerase. Annu. Rev. Microbiol 54, 499–518. [PubMed: 11018136]

Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, et al. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. Proc. Natl. Acad. Sci. USA 110, 13067–13072. [PubMed: 23878237]

Jeong Y, Kim J-N, Kim MW, Bucca G, Cho S, Yoon YJ, Kim B-G, Roe J-H, Kim SC, Smith CP, and Cho BK (2016). The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3(2). Nat. Commun 7, 11605. [PubMed: 27251447]

Johns NI, Gomes ALC, Yim SS, Yang A, Blazejewski T, Smillie CS, Smith MB, Alm EJ, Kosuri S, and Wang HH (2018). Metagenomic mining of regulatory elements enables programmable species-selective gene expression. Nat. Methods 15, 323–329. [PubMed: 30052624]

Kanehisa M, Sato Y, Kawashima M, Furumichi M, and Tanabe M (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44 (D1), D457–D462. [PubMed: 26476454]

Katoh K, Misawa K, Kuma K, and Miyata T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. [PubMed: 12136088]

Kelsic ED, Chung H, Cohen N, Park J, Wang HH, and Kishony R (2016). RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. Cell Syst. 3, 563–571.e6. [PubMed: 28009265]

Konaté MM, Plata G, Park J, Usmanova DR, Wang H, and Vitkup D (2019). Molecular function limits divergent protein evolution on planetary time-scales. eLife 8, e39705. [PubMed: 31532392]

Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. [PubMed: 19261174]

Letunic I, and Bork P (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47 (W1), W256–W259. [PubMed: 30931475]

Liu X, Brutlag DL, and Liu JS (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput 138, 127–138.

Love MI, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. [PubMed: 25516281]

Manganelli R, Provvedi R, Rodrigue S, Beaucher J, Gaudreau L, and Smith I (2004). σ factors and global gene regulation in Mycobacterium tuberculosis. J. Bacteriol 186, 895–902. [PubMed: 14761983]

McCutcheon JP, and Moran NA (2011). Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol 10, 13–26. [PubMed: 22064560]

Merrick MJ (1993). In a class of its own–the RNA polymerase sigma factor sigma 54 (sigma N). Mol. Microbiol 10, 903–909. [PubMed: 7934866]

Moran CP Jr., Lang N, LeGrice SFJ, Lee G, Stephens M, Sonenshein AL, Pero J, and Losick R (1982). Nucleotide sequences that signal the initiation of transcription and translation in Bacillus subtilis. Mol. Gen. Genet. 186, 339–346. [PubMed: 6181373]

Murakami KS, and Darst SA (2003). Bacterial RNA polymerases: the wholo story. Curr. Opin. Struct. Biol. 13, 31–39. [PubMed: 12581657]

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, and Bernardi G (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem. Biophys. Res. Commun. 347, 1–3. [PubMed: 16815305]

Muto A, and Osawa S (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA 84, 166–169. [PubMed: 3467347]

Paget MSB, and Helmann JD (2003). The sigma70 family of sigma factors. Genome Biol. 4, 203. [PubMed: 12540296]

Panaghie G, Aiyar SE, Bobb KL, Hayward RS, and de Haseth PL (2000). Aromatic amino acids in region 2.3 of Escherichia coli sigma 70 participate collectively in the formation of an RNA polymerase-promoter open complex. J. Mol. Biol 299, 1217–1230. [PubMed: 10873447]

Phillips R, Belliveau NM, Chure G, Garcia HG, Razo-Mejia M, and Scholes C (2019). Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression. Annu. Rev. Biophys 48, 121–163. [PubMed: 31084583]

Price MN, Dehal PS, and Arkin AP (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490. [PubMed: 20224823]

Raghavan R, Kelkar YD, and Ochman H (2012). A selective force favoring increased G+C content in bacterial genes. Proc. Natl. Acad. Sci. USA 109, 14504–14507. [PubMed: 22908296]

Rodrigue S, Provvedi R, Jacques PÉ, Gaudreau L, and Manganelli R (2006). The σ factors of Mycobacterium tuberculosis. FEMS Microbiol. Rev. 30, 926–941. [PubMed: 17064287]

Ronda C, Pedersen LE, Sommer MOA, and Nielsen AT (2016). CRMAGE: CRISPR Optimized MAGE Recombineering. Sci. Rep 6, 19452. [PubMed: 26797514]

Rosinski-Chupin I, Sauvage E, Sismeiro O, Villain A, Da Cunha V, Caliot ME, Dillies MA, Trieu-Cuot P, Bouloc P, Lartigue MF, and Glaser P (2015). Single nucleotide resolution RNA-seq uncovers new regulatory mechanisms in the opportunistic pathogen Streptococcus agalactiae. BMC Genomics 16, 419. [PubMed: 26024923]

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al. (2016). Local fitness landscape of the green fluorescent protein. Nature 533, 397–401. [PubMed: 27193686]

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 464, 250–255. [PubMed: 20164839]

Siegele DA, Hu JC, Walter WA, and Gross CA (1989). Altered promoter recognition by mutant forms of the σ 70 subunit of Escherichia coli RNA polymerase. J. Mol. Biol. 206, 591–603. [PubMed: 2661828]

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Sö ding J, et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol 7, 539. [PubMed: 21988835]

Starr TN, and Thornton JW (2016). Epistasis in protein evolution. Protein Sci. 25, 1204–1218. [PubMed: 26833806]

Stead MB, Agrawal A, Bowden KE, Nasir R, Mohanty BK, Meagher RB, and Kushner SR (2012). RNAsnap™: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. Nucleic Acids Res. 40, e156. [PubMed: 22821568]

Storz JF (2018). Compensatory mutations and epistasis for protein function. Curr. Opin. Struct. Biol 50, 18–25. [PubMed: 29100081]

Tomko TA, and Dunlop MJ (2017). Expression of Heterologous Sigma Factor Expands the Searchable Space for Biofuel Tolerance Mechanisms. ACS Synth. Biol. 6, 1343–1350. [PubMed: 28319371]

Urtecho G, Tripp AD, Insigne KD, Kim H, and Kosuri S (2019). Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. Biochemistry 58, 1539–1551. [PubMed: 29388765]

Waldburger C, Gardella T, Wong R, and Susskind MM (1990). Changes in conserved region 2 of Escherichia coli σ 70 affecting promoter recognition. J. Mol. Biol 215, 267–276. [PubMed: 2213883]

Wang HH, and Church GM (2011). Multiplexed genome engineering and genotyping methods: Applications for synthetic biology and metabolic engineering. Methods Enzymol. 498, 409–426. [PubMed: 21601688]

Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, and Church GM (2009). Programming cells by multiplex genome engineering and accelerated evolution. Nature 460, 894–898. [PubMed: 19633652]

Wong BG, Mancuso CP, Kiriakov S, Bashor CJ, and Khalil AS (2018). Precise, automated control of conditions for high-throughput growth of yeast and bacteria with eVOLVER. Nat. Biotechnol 36, 614–623. [PubMed: 29889214]

Zhang Y, Feng Y, Chatterjee S, Tuske S, Ho MX, Arnold E, and Ebright RH (2012). Structural basis of transcription initiation. Science 338, 1076–1080. [PubMed: 23086998]

Highlights

- Evolutionary sequence diversity of bacterial primary σ^{70} factors
- High-resolution fitness landscape map of *E. coli* σ^{70}
- Mapping natural diversity to fitness map predicts varying fitness deficits
- σ^{70} replacement in *E. coli* with natural orthologs elicits transcriptional changes

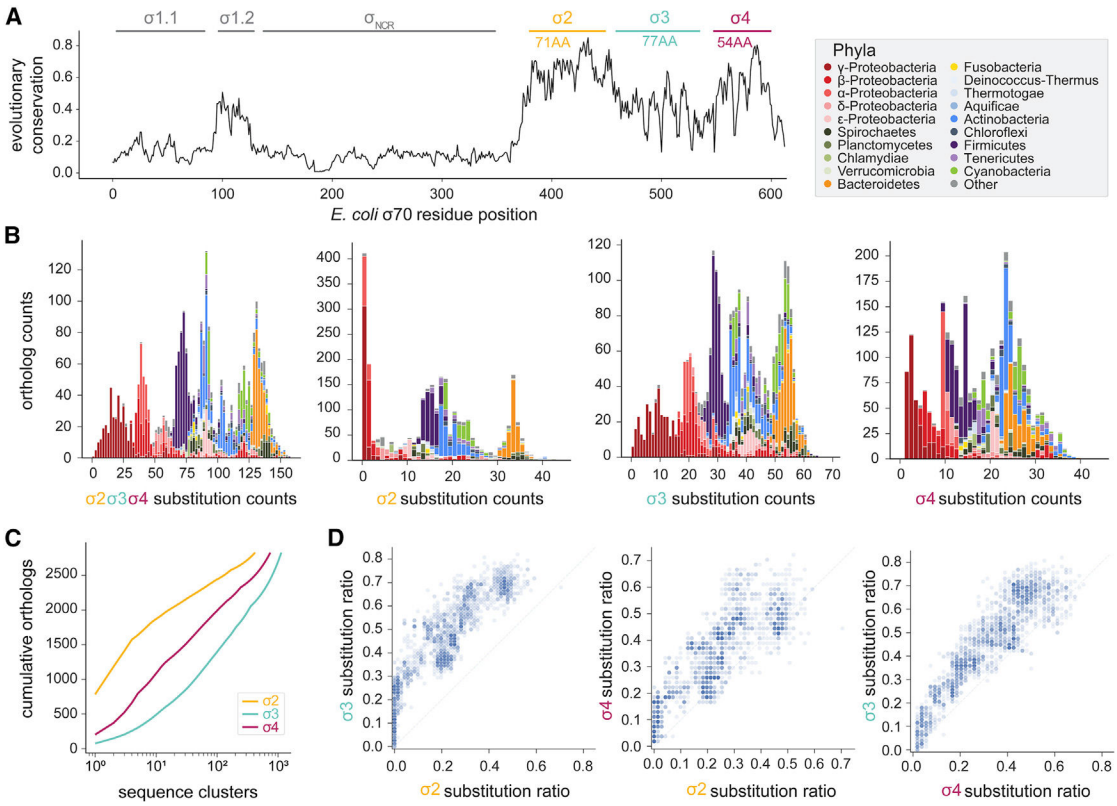


Figure 1. Evolutionary sequence analysis of primary σ^{70} orthologs

(A) Evolutionary conservation of primary σ^{70} (by Jensen-Shannon divergence) based on alignment of 4,702 σ^{70} sequences. Residue positions are based on *E. coli* σ^{70} with different domains shown (σ^{70}_1, σ^{70}_2, σ^{70}_3, and σ^{70}_4).

(B) Distribution of amino acid substitution counts of σ^{70} orthologs for domains 2–4 compared to the *E. coli* σ^{70} sequence. Colors in each bar correspond to ortholog host phylogeny at the phylum level, with the exception of Proteobacteria, which are separated at the class level.

(C) Cumulative distribution of orthologs clustered to 90% sequence identity for each domain.

(D) Pairwise comparisons of substitution count ratios between σ^{70} domains. Dashed lines denote 1:1 ratio.

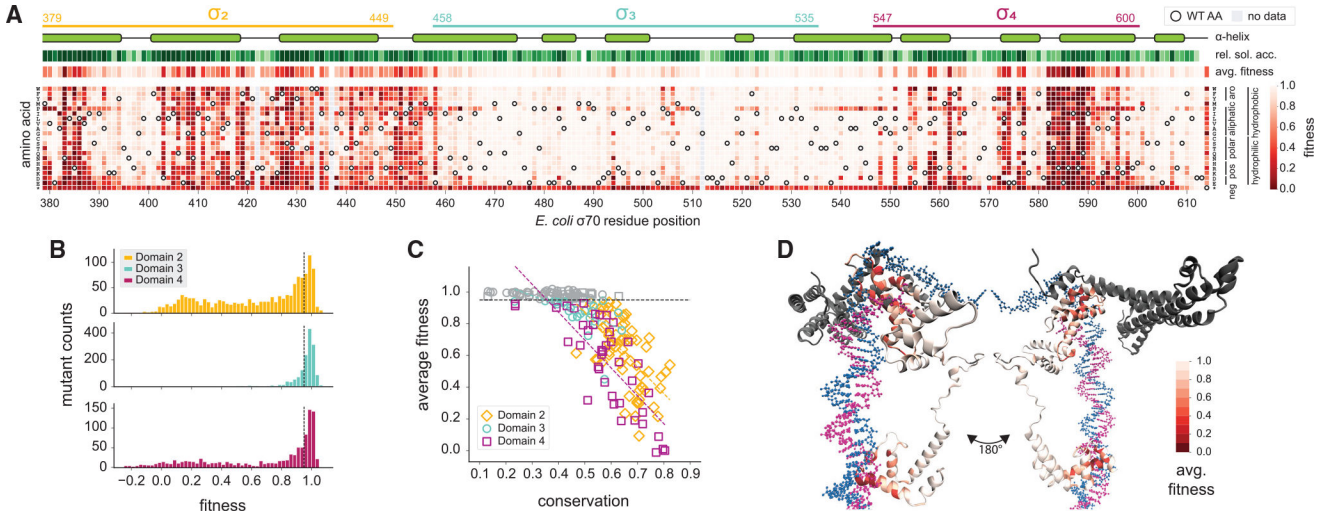


Figure 2. Mapping the fitness landscape of *E. coli* σ^{70}

(A) Fitness landscape of *E. coli* σ^{70} at residues 379–613 profiled by MAGE-seq. Columns of the heatmap correspond to positions along the σ^{70} protein and rows correspond to all 20 amino acid residues plus stop codons (*). Open circles denote the wild-type *E. coli* σ^{70} residue at each position. Gray squares denote data not available. Regions of structured alpha helices, relative solvent accessibility, and average fitness at each residue position are displayed above the heatmap.

(B) Histogram of the distribution of fitness effects (DFEs) for each σ^{70} domain. Dotted lines denote fitness of 0.95, deemed as the separation between neutral and detrimental fitness.

(C) Scatterplot of σ^{70} evolutionary conservation and mean fitness for each residue position. Neutral residues (fitness ≥ 0.95) are displayed in gray, while detrimental residues (fitness < 0.95) are colored by their respective domains. Colored dash lines indicate linear regressions of detrimental residue positions in each domain.

(D) Protein structure of σ^{70} (ribbon model) bound to an open DNA complex (stick model) using PDB: 6CA0. Red color scale represents mean fitness at each residue position on the σ^{70} structure; dark gray regions are residues not profiled with MAGE-seq.

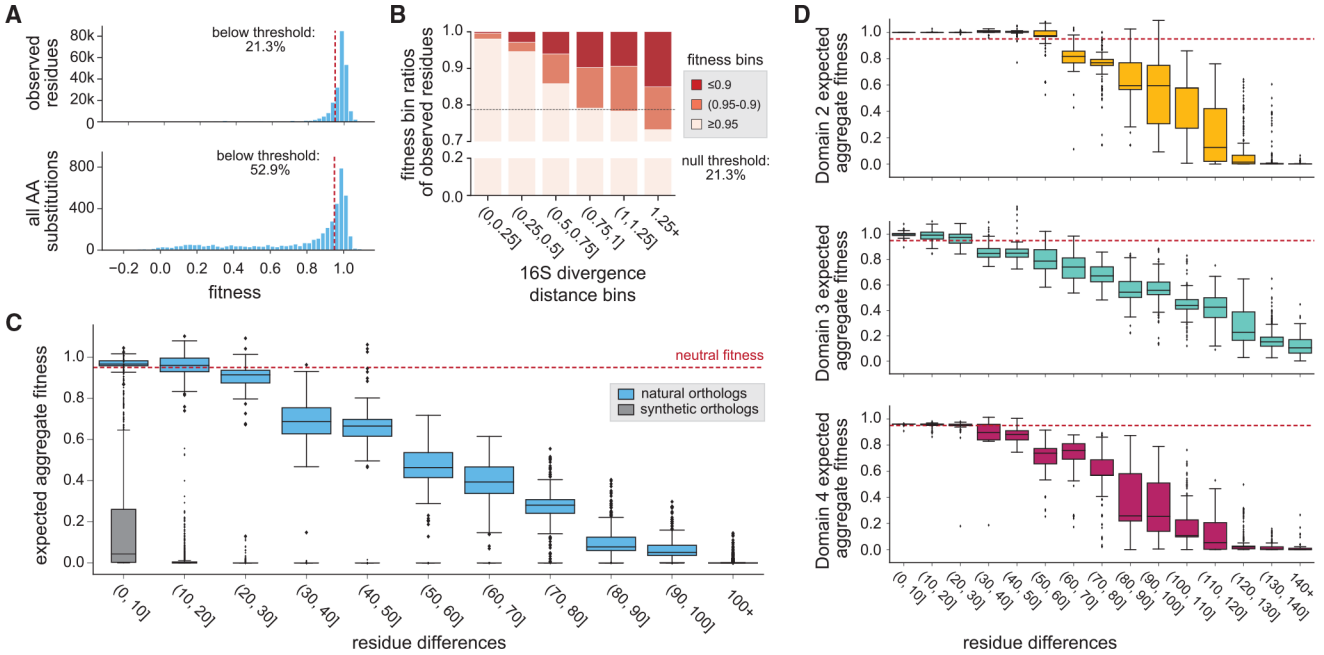


Figure 3. Fitness predictions of orthologous σ^{70} sequences using *E. coli* σ^{70} fitness landscape

(A) DFEs of residue variants observed in natural ortholog sequences (top) compared to DFEs of all possible single-residue mutations in σ^{70} (bottom). Fitness threshold of 0.95 is designated by the dotted line.

(B) Plot of residue fitness in σ^{70} orthologs versus binned 16S phylogenetic distance to *E. coli* showing higher fraction of deleterious fitness variants at greater evolutionarily distance from *E. coli*.

(C) Blue boxplots show expected aggregate fitness (EAF) distributions of natural orthologs (pink) with increasing binned number of residue differences to *E. coli* σ^{70}. Fitness at 0.95 is denoted by the dotted line. Gray boxplots show null EAF distributions of synthetically generated σ^{70} sequences with random mutations at each residue difference bin.

(D) EAF for each σ^{70} domain against the total binned number of residue differences across domains 2–4.

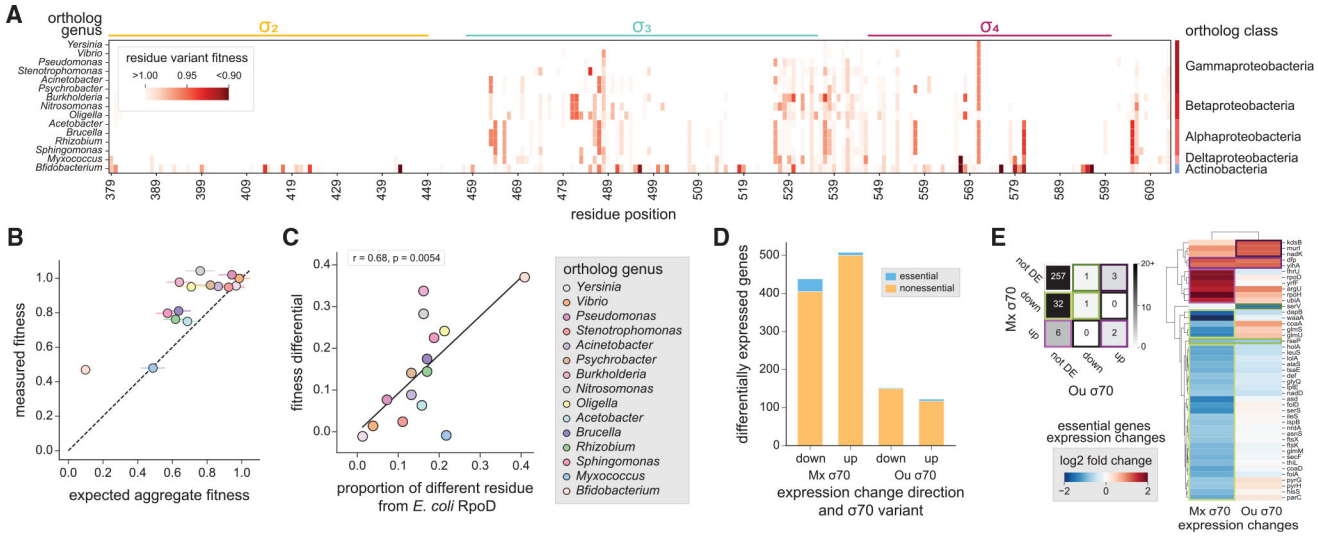


Figure 4. Characterization of orthologous σ^{70} sequence variants in *E. coli*

(A) Residue-level fitness map in the 15 σ^{70} orthologs measured in *E. coli*.

(B) Plot of EAF of each σ^{70} ortholog and their measured fitness in *E. coli*. Measured fitness values represent the average from two independent fitness competition experiments.

(C) Plot of fitness differential (measured fitness minus EAF) and proportion of residue differences of orthologs from *E. coli* RpoD.

(D) Number of differentially expressed genes in Mx σ^{70} and Ou σ^{70} transcriptomes compared to Ec σ^{70}.

(E) The number and grouping of essential genes that are differentially upregulated or downregulated in Mx σ^{70} and Ou σ^{70}.

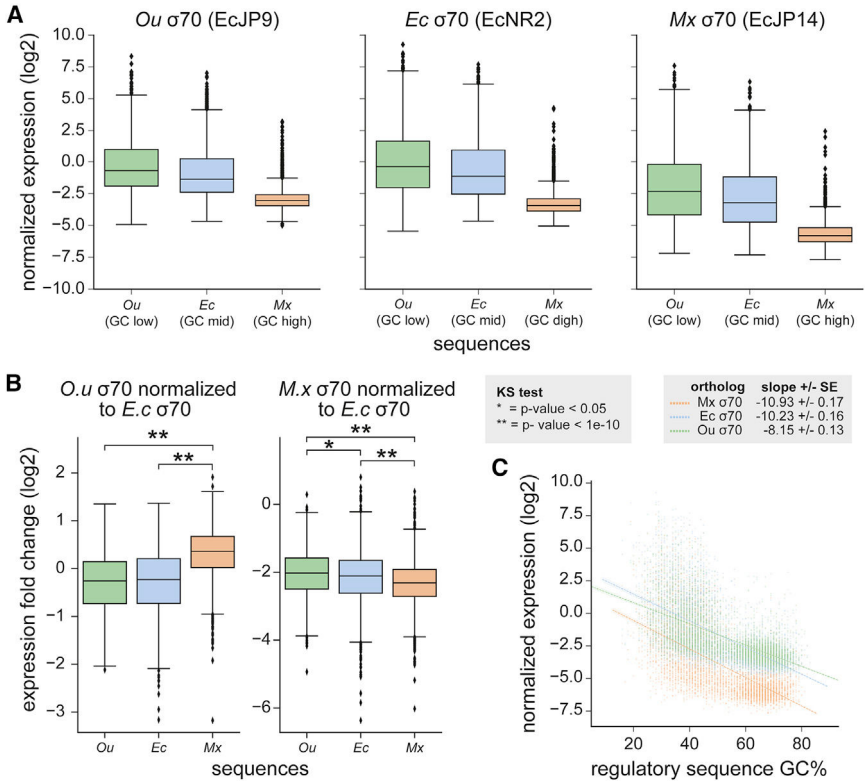


Figure 5. Multiplexed transcriptional measurements of a metagenomic regulatory sequence library in *E. coli* strains with orthologous σ^{70}

(A) Distribution of normalized Tx values of regulatory sequences transcribed by strains expressing Ou σ^{70} (EcJP9), Mx σ^{70} (EcJP14), and Ec σ^{70}. Regulatory sequences grouped by their genomic origins is shown in the boxplots in each graph.

(B) Expression fold change by Ou σ^{70} and Mx σ^{70} normalized to Ec σ^{70} data (statistical significance based on KS test; p values shown in key).

(C) Relationship between GC content and Tx values of regulatory sequences using different σ^{70} proteins. Linear regression for each dataset is plotted.

Table 1.

E. coli strains with σ^{70} orthologs

Strain	Class	Genus	Residue differences	Expected aggregate fitness	Measured fitness
EcJPI	Gammaproteobacteria	*Yersinia*	3	0.968	0.957
EcJP2	Gammaproteobacteria	*Vibrio*	9	0.985	0.999
EcJP3	Gammaproteobacteria	*Pseudomonas*	17	0.943	1.020
EcJP4	Gammaproteobacteria	*Stenotrophomonas*	26	0.924	0.948
EcJP5	Gammaproteobacteria	*Acinetobacter*	31	0.864	0.953
EcJP6	Gammaproteobacteria	*Psychrobacter*	31	0.820	0.961
EcJP7	Gammaproteobacteria	*Burkholderia*	38	0.640	0.978
EcJP8	Betaproteobacteria	*Nitrosomonas*	38	0.760	1.043
EcJP9	Betaproteobacteria	*Oligella*	50	0.709	0.950
EcJP10	Betaproteobacteria	*Acetobacter*	37	0.687	0.750
EcJPI1	Alphaproteobacteria	*Brucella*	40	0.636	0.811
EcJP12	Alphaproteobacteria	*Rhizobium*	40	0.618	0.762
EcJP13	Alphaproteobacteria	*Sphingomonas*	44	0.572	0.797
EcJP14	Deltaproteobacteria	*Myxococcus*	51	0.488	0.479
EcJP15	Actinobacteria	*Bifidobacterium*	96	0.099	0.469

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
E. coli ECNR2	Addgene / Wang et al., 2009	26931
RpoD ortholog mutants	This paper	See Table S1
Chemicals, peptides, and recombinant proteins		
Kanamycin	Fisher Scientific	BP9065
anhydrous tetracycline	Cayman	10009542
Critical commercial assays		
Genomic DNA prep kit	GE	28904259
Q5 polymerase	NEB	M0543L
Maxima H minus reverse transcriptase	Thermo Scientific	EO0751
Sybr Green	Invitrogen	S7567
SPRI beads	Beckman Coulter	A63881
NextSeq 300-cycle kit	Illumina	20024905
Miseq 300-cycle kit	Illumina	MS-103–1002
Deposited data		
RpoD ortholog mutants RNA-seq raw data	ArrayExpress	E-MTAB-9099
MAGE-seq raw data	ArrayExpress	E-MTAB-9103
Metagenomic regulatory sequence library raw data	ArrayExpress	E-MTAB-9111
Oligonucleotides		
MAGE oligos	This study	See Data S2
Oligos used in this study	This study	See Data S2
Regulatory sequence oligo library	Agilent/This study	G7721A/See Data S5
Recombinant DNA		
Orthologous sigma factor sequences	This study	See Table S2
Software and algorithms		
Clustal Omega	Sievers et al., 2011	NA
FastTree2	Price et al., 2010	NA
iTOL	Letunic and Bork, 2019	NA
BioProspector	Liu et al., 2001	NA
Trimmomatic	Bolger et al., 2014	NA
Bowtie	Langmead et al., 2009	NA
HTSeq	Anders et al., 2015	NA
DEseq2	Love et al., 2014	NA

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MAGE-seq analysis code files	This study	https://github.com/jiminpark66/MAGEseq
Regulatory sequence library analysis code files	This study	https://github.com/jiminpark66/regulatorysequence_library
Other		
Turbidostat	This Study	NA
eVOLVER	Wong et al., 2018	NA