HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 September 19.

Published in final edited form as:

Nat Methods. 2018 May ; 15(5): 323–329. doi:10.1038/nmeth.4633.

Metagenomic mining of regulatory elements enables programmable species-selective gene expression

Nathan I. Johns[1,2,#], **Antonio L.C. Gomes**[1,#,†], **Sung Sun Yim**[1], **Anthony Yang**[4], **Tomasz Blazejewski**[1,2], **Christopher S. Smillie**[5], **Mark B. Smith**[6], **Eric J. Alm**[5,6,7,8], **Sriram Kosuri**[9,10,11], and **Harris H. Wang**[1,3,*]

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

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

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

[4]School of Engineering and Applied Sciences, Columbia University, New York, NY, USA

[5]Broad Institute, Cambridge, MA, USA

[6]Department of Biological Engineering, MIT, Cambridge, MA, USA

[7]Computational and Systems Biology Initiative, MIT, Cambridge, MA, USA

[8]The Center for Microbiome Informatics and Therapeutics, MIT, Cambridge, MA, USA

[9]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA, USA

[10]UCLA-DOE Institute for Genomics and Proteomics, University of California, Los Angeles, Los Angeles, CA, USA

[11]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA, USA

Abstract

Robust and predictably performing synthetic circuits rely on the use of well-characterized regulatory parts across different genetic backgrounds and environmental contexts. Here, we report the large-scale metagenomic mining of thousands of natural 5′-regulatory sequences from diverse bacteria and their multiplexed gene expression characterization in industrially-relevant microbes. We identified regulatory sequences with broad and host-specific expression properties that are

*Correspondences should be addressed to H.H.W. (hw2429@columbia.edu).

#These authors contributed equally.

†Present Address: Department of Immunology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

AUTHOR CONTRIBUTIONS

N.I.J., A.L.G., C.S.S., M.B.S., E.J.A, S.K., and H.H.W. designed the study. N.I.J., S.S.Y., and H.H.W. performed the experiments. N.I.J, A.L.G., A.Y., T.B., and H.H.W. analyzed the data. N.I.J, A.L.G., and H.H.W. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

robust in various growth conditions. We further observed significant differences between species' capacity to utilize exogenous regulatory sequences. Finally, we demonstrated programmable species-selective gene expression that produces distinct and diverse output patterns in different microbes by leveraging regulatory sequences with pre-defined host-specificities. Together, these findings provide a rich resource of characterized and annotated natural regulatory sequences and a framework to engineer synthetic gene circuits with unique and tunable cross-species functionality and properties.

Introduction

Synthetic biology relies on well-characterized genetic components or parts to modularly assemble increasingly sophisticated gene circuits with specified function[1]. Recent advances in high-throughput DNA sequencing and synthesis have greatly increased our ability to generate new genetic parts[2]. Natural enzymes and regulatory proteins have been systematically screened for new functionality[3-5], while non-coding cis-regulatory elements have been characterized to better understand biophysical parameters[6], parts composability[7], contextual robustness[8] and regulatory logic[9] for building more complex genetic systems. Most regulatory components are derived from mutational variants templated from a few sequences of limited genetic diversity[10,11]. The vast majority of parts used today are based on those from a few model organisms[12] and their functionality in diverse genetic backgrounds and growth conditions remain poorly characterized. For many commercially useful microbes, only a handful of regulatory parts have been rigorously tested, and these often have limited range of expression[13-18]. Efforts to utilize exogenous regulatory parts in new hosts often fail due to differences in their gene expression machinery[19]. While more universally compatible and portable regulatory systems have been proposed using orthogonal regulators[5,20-22], these approaches still rely on endogenous machineries for initial activation, which are uncharacterized for most species. The development of regulatory parts with programmable host-ranges could enable new types of synthetic circuits for engineering diverse microbial communities for industrial and therapeutic applications[23].

Here, we report the mining of 184 microbial genomes to yield a diverse library of tens of thousands of natural regulatory sequences. We systematically quantified transcription and translation levels of the library across different bacterial species and growth conditions and developed species-selective gene circuits with distinct preprogrammed output patterns in different hosts. This dataset significantly expands the repertoire of prokaryotic regulatory sequences that can be used to build synthetic circuits with new layers of sophistication in multi-species bacterial communities.

RESULTS

Mining and characterizing natural regulatory sequences

To expand the phylogenetic breadth of useful regulatory components (i.e. promoters, translation initiation signals), we first mined 184 prokaryotic genomes for putative regulatory sequences (Figure 1, *See Methods*). These genomes spanned major phylogenetic groups from diverse habitats and included industrially relevant species (Suppl. Figures S1,

Suppl. Data Table 1). We compiled a library of 29,249 uniquely barcoded regulatory sequences (RSs) with an average of 159 derived from each genome.

To determine the transcription and translation activity of each regulatory sequence in the library, we utilized a previously described high-throughput GFP reporter system[7] (Figure 1). The RS library was generated by microarray oligo synthesis, amplified, and cloned as a pool into shuttle vectors (Suppl. Figure S2) upstream of a super-folding GFP and subsequently transformed into different species for characterization. To determine transcription levels of the RS library, we used targeted RNA-seq and DNA -seq and normalized each construct's sfGFP mRNA read counts by its total DNA abundance in the population after filtering for sequencing and synthesis errors. These multiplex transcription measurements showed high degrees of concordance between biological replicates and duplicate RSs with alternate barcodes (Pearson $r = 0.88$ and 0.86 respectively, Suppl. Figure S3). RT-PCR measurements of individual library members are also highly correlated with their corresponding multiplex measurements (Suppl. Figure S4a). To measure translational activity, we employed FACS-seq to quantify sfGFP protein levels generated from each regulatory sequence (Suppl. Figure S4b)[6,7]. Flow cytometry measurements of isolate library members showed a high level of correlation with their population-derived FACS-seq library data (Suppl. Figure S4c). Furthermore, transcription and translation measurements using an alternative reporter, mCherry, were well-correlated with GFP values (Suppl. Figure S5).

Universal and host-specific patterns of transcriptional activation

To explore the transcriptional potential of our RS library in different bacterial hosts, we first transformed the library at high coverage into *Bacillus subtilis*, *Escherichia coli*, and *Pseudomonas aeruginosa*. *B. subtilis* is a soil gram-positive Firmicute, while *E. coli* and *P. aeruginosa* are gram-negative Proteobacteria that colonize diverse environments. Transcriptional measurements were made from mid-exponential phase cultures, resulting in a converged set of 11,319 regulatory constructs with high-confidence expression across each species. To enable transcription profile comparisons between species, we normalized transcription values in each species with endogenous control sequences present in the library, which are used as references to compare RS library activity levels with sequences that are representative of each host's native transcriptome (*See methods*).

Significant differences in the transcription activity of the RS library between different hosts were observed (Figure 2a, Suppl. Data Table 2). *B. subtilis* displayed the lowest number of measurably active RSs (18.9% with > 0 RNA reads), while *E. coli* and *P. aeruginosa* had substantially higher fractions of active RSs with measureable transcription activity (52.0% and 83.8% respectively). In each species, expression levels spanned several orders of magnitude, indicating diverse transcriptional functionality across the library. Comparison of these expression profiles between species revealed four general groups: universally active (16.9%), differentially active in 2 of 3 species (33.3%), specifically active in 1 species (37.4%), or inactive in all 3 species (12.4%). In general, universally active regulatory sequences had lower GC content than the overall library (Figure 2b). Interestingly the converse is observed on the host side, with each organism's capacity to utilize exogenous regulatory sequences appearing to correspond with increasing genomic GC content – *P.*

aeruginosa (66% GC) activated the largest fraction RSs, followed *by E. coli* (50% GC) and *B. subtilis* (42% GC).

While closely related species might be expected to have regulatory systems that are more cross-compatible, this has not been systematically studied. We filtered the RS library phylogenetically for only donor sequences from *Bacillaceae, Enterobacteracea*, or *Pseudomonaceae* families and analyzed their activity in the three recipients. Interestingly, we identified distinct patterns of intra- versus inter-family transcriptional specificities (Figure 2c). *B. subtilis* could activate 47.7% of donor *Bacillaceae* regulatory sequences, but only 10.8% of *Enterobacteracea* and 3.2% of *Pseudomonaceae* RSs. *E. coli* and *P. aeruginosa* were better able to express foreign regulatory sequences, with each activating a larger fraction of all three donor RS families. Mined *Bacillaceae* sequences showed more broad range activity (>45% of sequences) in all three recipients and a higher mean expression level especially in non-*Bacillaceae* recipients (Figure 2c). In contrast, *Pseudomonaceae* sequences were generally not expressed in *B. subtilis* or only expressed at low levels in *E. coli*, highlighting stringent host-specificity of its regulatory signals.

We further delineated the regulatory architecture of each sequence by identifying transcription start sites (TSSs) based on our targeted RNA-seq reads. Most TSSs fell between −20 and −50 bp from the start codon (Suppl. Figure S6), consistent with known native promoter architectures in many bacteria[24-26]. This dataset should improve efforts to model bacterial transcription and design new gene circuits. Together, these results highlight that prokaryotic genomes are a rich reservoir for mining functional regulatory parts with diverse cross-species properties that can be systematically quantified using high-throughput library synthesis and transcriptional profiling.

Since environmental and growth conditions induce changes in gene expression, we also explored the extent to which the RS library activity is dependent on growth phase or environmental conditions experienced by the host. We measured RS library transcriptional activity in *E. coli* under five different growth and stress conditions (Suppl. Figure S7, Suppl. Data Table 3). Many regulatory sequences (17.3%) exhibited universally high activity across all conditions, while others showed differentially moderate to low transcription activity (28.6 and 22.8% respectively). TSSs tended to be highly conserved across growth conditions (Suppl. Figures S7-S8). To generate a set of regulatory sequences with robust untranslated regions (UTRs) and transcriptional activities across growth conditions in *E. coli*, we further filtered the RS library down to a list of 100 sequences with a wide range of transcription activity from only a single TSS (Suppl. Figure S7d). We expect this robust RS sub-library to be a useful resource for circuit designs that will be deployed in diverse environments. The use of diverse sequences may also improve DNA assembly efficiencies of larger and more complex gene circuits[27] as well as better maintain their evolutionary stability[28].

Predictive features of transcriptional activity

To determine regulatory sequence features that govern transcription levels, we performed *de novo* motif finding using MEME[29]. For each host, the promoter library was divided into four groups based on their activity levels (Suppl. Figure S9a). A common motif was enriched in high activity promoters in all recipients, which corresponded to the canonical binding motif

for the housekeeping σ70 factor (Figure 3a). Searches for additional motifs yielded only degenerate versions of the core σ70 motif (Suppl. Figure S9b,c).

To develop a predictive model of transcription activity, we investigated three factors that could influence gene expression, promoter GC content, σ70 binding affinity and 5′ mRNA stability. Promoter GC content indicates compositional preferences of sequence elements that could promote transcription. σ70 is the dominant and most abundant σ-factor and is responsible for transcription of a wide array of housekeeping genes[30,31]. Secondary structure of mRNA affects the rate of mRNA decay[32,33], which combined with the transcription rate determines overall mRNA transcript levels. Each of the parameters are correlated with measured transcription activity of the regulatory sequence library (Figure 3b). Higher promoter GC content is anti-correlated with transcription activity, while a match to the σ70 binding motif is positively-correlated with activity, as is lower RNA stability (i.e. higher ΔG folding energy). Controlling for these parameters independently showed that the σ70 binding motif is most informative for assessing transcription activity (Suppl. Figure S10). Integration of these parameters into a linear regression model showed a predictive power of 32%, 69%, and 54% for the variances of transcription activity in *B. subtilis*, *E. coli* and *P. aeruginosa*, respectively (Figure 3c). These results demonstrate that a simple model can explain a significant fraction of the variation observed in transcriptional activity within different hosts.

Translational activity of RS library across hosts

While transcriptional activation in bacteria is mediated by transcription and sigma factor recruitment of the RNA polymerase complex, translation initiation is mediated by interactions between ribosomal subunits and the mRNA transcript. *In silico* modeling of factors that govern ribosomal initiation have enabled the generation of predictive algorithms for bacterial translation rates[28]. However, the cross-compatibility of translation initiation sequences from different species has not been characterized. To tackle this challenge, we systematically quantified the amount of fluorescence generated from each regulatory sequence in our library in high-throughput across three recipients using FACS-Seq (Figure 4a)[6,7]. Across the recipients, we identified a shared set of 8,898 regulatory sequences that spanned nearly 3 orders of magnitude of fluorescence (Suppl. Figure S11a), with 3.3% of the library (290 constructs) expressing GFP proteins in all species (Suppl. Figure S11b). Examination of sequences in the region upstream of highly translated library members revealed enrichment of A and G bases centered near 10 bp upstream from the start codon (Suppl. Figure S11c).

To probe the differential impact of transcription and translation requirements for gene expression across recipients and for different donor groups, we stratified the regulatory activation profile of the RS library across bins of transcription and translation levels (Figure 4b). Overall, higher transcriptional activity is associated with higher GFP levels, although translation rates varied widely even for highly transcribed regulatory sequences. Normalization over transcription or translation bins highlighted distinct patterns of regulatory specificities associated with RNA or protein generation. Regulatory sequences belonging to low transcription bins generally do not yield GFP signal, indicating that

transcription is a key barrier in gene expression in these cases. While *P. aeruginosa* was able to transcribe a large fraction of the RS library (83%), only 9% of those RNA species ultimately yielded significant GFP fluorescence, which may reflect incompatibilities at the level of translation (Figure 4c). In contrast, of actively transcribed sequences, *B. subtilis* and *E. coli* were able to yield significant GFP levels in 20-30% of these RNA transcripts. Interestingly, RSs from Firmicutes species showed a high potential to be both transcribed and translated in each host organism (Suppl. Figure S12a). In contrast, while RSs from Proteobacteria species could be transcribed and translated in *E. coli* and *P. aeruginosa*, they were often either not transcriptionally active in *B. subtilis* or further translationally limited even for transcribed RNAs (Suppl. Figure S12b). We additionally assessed the transcription activity and translation efficiency of 212 RSs that contained both active transcription and translation data across all species (Suppl. Figure S13a). Translation efficiency of each RS was determined by normalizing its GFP level to its transcription level. Interestingly, we find that between recipients, only *E. coli* and *P. aeruginosa* showed significant correlations between regulatory sequences in terms of transcription levels and translation efficiencies. Finally, we predicted translation initiation efficiency of UTRs generated from each regulatory sequence using the RBS calculator v1.0[34] and found reasonable correlation between predicted values and experimental data (Suppl. Figure 13b).

Together, these results highlight that even if there are similar regulatory specificities at the transcription and translation levels between two species, both processes play distinct roles in functionalizing heterologous regulatory sequences with possible separate barriers to expression. Moreover, some species (e.g. *B. subtilis*) naturally possess highly restrictive transcriptional and/or translational requirements for gene expression, which suggests the possibility that these differential specificities across hosts could be exploited as pre-defined parameters to design genetic circuits for deployment in multi-species microbial communities.

Expanding RS library characterization to other hosts

To further extend the characterization of the RS library, we selected 241 library members (RS241 library), cloned and introduced them into additional industrially useful hosts *Salmonella enterica*[35], *Vibrio natriegens*[36] (both Gammaproteobacteria) and *Corynebacterium glutamicum*[37] (a gram-positive Actinobacteria). Multiplex measurements of RS241 in *B. subtilis*, *E. coli*, *P. aeruginosa* and these three new hosts exhibited activity spanned nearly 6 orders of magnitude for transcription and 3 orders of magnitude for translation (Suppl. Figures S14, Suppl. Data Table 4). We observed differential compatibility of regulatory sequence performance for transcription and translation across phylogenetically diverse species (Suppl. Figure 15). These results highlight the utility of multiplexed measurements of small targeted libraries among organisms where large scale characterization may be challenging.

Programming species-selective gene expression patterns

Engineering host-specific regulation enables the development of cross-species genetic programs that generate complex behavior in mixed communities. For example, a broad host-range transmissible plasmid can be designed to generate different pre-defined behaviors

from the same DNA sequence depending on specificity to the host regulatory machinery (e.g. activation of function only in subset of species). Targeting sub-populations in a mixed consortium constitutes a powerful strategy for community-level microbiome engineering[38-40]. We explored the development of programmable "Species-selective Gene Circuits" (SsGC) that exploit natural host-specificity of heterologous regulatory sequences in different bacteria. By leveraging the universal and orthogonal regulatory activation properties exhibited in our RS library, we built simple dual-reporter that produced distinct fluorescence states depending on the recipient-context (Figure 5a).

We paired 12 regulatory sequences to drive a dual mCherry-GFP reporter construct in a broad host-range vector pNJ6.2, with each regulator independently controlling each fluorescent protein. Each construct was introduced into three recipients (*B. subtilis*, *E. coli*, and *P. aeruginosa*) to characterize their host-dependent behaviors. Across 10 SsGC constructs (A-J), we demonstrated distinct states of the two reporters ranging from universal, host-specific, and host-excluding activation profiles across recipients (Figure 5b). Some SsGCs exhibited universal activation across all hosts in both reporters (constructs A-C) while others had universal activation for mCherry but not sfGFP for *B. subtilis* (constructs D-E). We also built SsGCs that demonstrated the ability to selectively exclude expression of one fluorescence protein in *E. coli* only but not the other species for both reporters (constructs F-I). Additionally, we showed a SsGC exhibiting universal activation of GFP while mCherry expression was limited only to *P. aeruginosa* (construct J), demonstrating the possibility to specifically express one gene in only a single defined species while other components are expressed more broadly across multiple species. These designs constitute a first step towards generating more complex functions that could be differentially activated across multiple species of a diverse microbial community towards engineering sophisticated community-level dynamics and behaviors.

DISCUSSION

Characterizing regulatory part performance across different host organisms and growth conditions is crucial for programming gene circuits of increasing sophistication and reliability. Here, we combined metagenomic mining, oligo library synthesis, and high-throughput characterization to measure transcriptional and translational activities of tens of thousands of natural regulatory sequences across up to six diverse bacterial species and under multiple growth conditions. We find substantial differences in the ability of each species to transcribe and translate exogenous regulatory sequences. For instance, *P. aeruginosa* was able to activate the largest fraction of the library we tested, followed by *E. coli*, and *B. subtilis*. *B. subtilis* showed extremely limited regulatory activation potential—a pattern that appears to be associated with the host species' genomic GC content. We speculate that evolution towards different genomic GC contents may influence the capacity of gene expression machineries to utilize regulatory elements of varying sequence compositions. Importantly, we identified and annotated regulatory sequences with both universal and orthogonal host-ranges, which represent a rich resource for synthetic biology applications that rely on well-characterized components across different host backgrounds. Characterization of a subset of the RS library in *C. glutamicum, V. natriegens*, and *S. enterica* further enhances the utility of this resource for tuning gene expression across a wide

range of activity levels in industrially relevant bacteria using a common set of regulatory sequences.

To demonstrate the application of these universal and host-specific regulatory sequences, we built simple species-selective dual-reporters that have defined activity profiles across three bacterial species. We successfully demonstrated circuits where two proteins have independent host expression profiles of varying specificity. These demonstrations are a first step towards designing more complex cross-species constructs that exhibit pre-defined behaviors depending on the host species. Functionalizing gene circuits to specific species is a useful strategy for microbiome perturbations (e.g. deploying biosensors in specific species[41] or eradicating pathogenic strains[38,39] by targeted toxin expression). With further advances in gene delivery technologies for *in situ* microbiome engineering[23], we expect strategies that leverage host regulatory differences will play a key role in controlling and maintaining synthetic circuit function and performance, especially when circuits can propagate in multiple hosts yet only activate in specified species.

ONLINE METHODS

Bacterial strains and expression vector construction

E. coli MegaX DH10B Electrocomp cells (ThermoFisher C640003) were used for all initial library cloning steps. Recipient test strains were *Escherichia coli* MG1655, *Bacillus subtilis* BD3182 (a 168 type strain derivative with Δ*rok*::kanR, Met⁻, Leu⁻, His⁻ to improve transformation; courtesy of D. Dubnau), and *Pseudomonas aeruginosa* PAO1 (with Δ*psy2* to remove pyocin S2 autofluorescence; courtesy of A. Rasouly and S. Lory). *V. natriegens* 14048, *C. glutamicum* 13032, and *S. enterica* Typhi Ty2 were obtained from ATCC.

Separate reporter plasmids were designed and constructed for each species pNJ1, pNJ2.1, pNJ3.1 using the backbones pZA11 (p15A ori, 11 copies / cell), pDG1662 (integration into *amyE* locus)[42], and pJN105 (pBBR1 ori, 20 copies / cell)[43] respectively. Unwanted restriction sites for PstI, EcoRI, and BamHI found outside of multi-cloning sites were removed by isothermal assembly. An ATG-less sfGFP construct[44] with upstream 5′ BamHI, spacer, PstI and downstream EcoRI was then cloned into each backbone in order to create the final reporter plasmids (Suppl. Figure S3). The broad-host vector pNJ6.2 was generated by first introducing the entire amyE-L to amyE-R region of pNJ2.1 into pNJ3.1. Subsequently, a reverse direction mCherry gene was placed just upstream of the amyE-L arm (see Figure 5a). For small library experiments, pNJ7 and pNJ8 were constructed from plasmids pACYC184 and pCES208 for *V. natriegens* and *C. glutamicum* respectively.

Metagenomic regulatory sequence library design

The 184 annotated and complete genomes were chosen from the Integrated Microbial Genomes Database[45] to maximize representation of microbes across the tree of life and to include industrially or medically relevant representative species, which included 169 bacteria and 15 archaea. For each genome, we identified all unidirectional intergenic regions (i.e. preceding and following genes on the same strand to avoid bidirectional elements) greater than 200 bp in size and extracted the 165 bp immediately upstream of annotated start

codons. These sequences will be referred to as RSs for convenience. RSs containing BamHI, PstI, and EcoRI sites were filtered out. We randomly chose subsets of RSs from each species, yielding ~160 sequences per genome (Suppl. Figure S1), which totaled a final library of 29,249 RSs. For each RS, we noted the COG category of the downstream gene being regulated, although no bias was introduced during random sub-selection of the RS sequences. We then added BamHI and PstI cut sites, a start codon, a unique 12 bp barcode (Levenshtein distance of >2), and common amplification sequences to the RSs as shown in Figure 1. We randomly selected a subset of 4,778 RSs from the total library to encode a different set of 12 bp barcodes as an internal control to assess the impact of barcode sequences on gene expression. In total, a 230 bp oligo pool containing 34,027 RSs was synthesized.

Library synthesis, cloning, and transformation into diverse hosts

All enzymes were obtained from New England Biolabs unless specified otherwise. The metagenomic RS library was synthesized as a 1 pmol oligo mix by Agilent Technologies (Carlsbad, CA) using their oligo library synthesis (OLS) platform[46]. The oligo library was first amplified for 8 cycles to make a template stock (amp1). All subsequent amplifications used this template as input DNA to avoid freeze-thaw cycles of the original oligo library stock. We performed a second amplification step using 1 uL of purified amp1 template stock to obtain enough DNA of the library (amp2) for cloning by performing 8 parallel qPCR reactions that were stopped after the reaction exited exponential amplification phase (usually ~8-10 cycles). All reactions used Kapa SYBR Fast Mastermix and were performed on a CFX96 Touch Real-Time PCR machine (Bio-Rad). Amplified library DNA was purified, digested with BamHI and PstI and ligated into each plasmid backbone using T4 DNA Ligase. Ligations were transformed into *E. coli* MegaX DH10B electrocompetent cells (Life Technologies). A 10 μL aliquot of each electroporation recovery mixture was diluted and plated to determine the cloning efficiency and library coverage, while the remaining 990 uL was propagated through two subsequent liquid selections in 25 mL LB-Lennox (BD Biosciences)+50 μg/mL carbenicillin grown at 30°C, 250 rpm overnight. All libraries were cloned with >50× coverage as determined by dividing the number of CFUs by the size of the designed library. Plasmid DNA was then extracted from library cultures using a Qiagen Midiprep kit for subsequent transformation into final the host strains.

Plasmid libraries were transformed into electrocompetent *E. coli* MG1655 by pelleting and washing a 100 mL mid-log phase culture with 10% glycerol at 4 °C three times and suspending the final pellet in 100 μL. Plasmid library DNA (1ul, 50-100ng) was added to multiple 20 μL aliquots of competent cells and electroporated at 1.8kV using a BioRad Micropulser. The cultures were recovered in 1 mL SOC for 1 hour at 30 °C, 250 rpm. We determined the library coverage by plating up to 1% of the transformed population on selective plates. The remaining 99% of the transformation culture post-1hr recovery was passaged through two subsequent liquid selections in 25 mL LB-Lennox+50 μg/mL carbenicillin grown at 30 °C, 250 rpm overnight to yield the final *E. coli* RS library.

B. subtilis BD3182 was transformed by diluting an overnight culture 1:100 into competence media containing 1× Spizizen salts supplemented with 0.5% glucose, 0.02% casein

hydrolysate, 0.1% yeast extract, 2.5 mM $MgCl_2$ and 50 ug/mL of histidine, leucine, and methionine. The culture was grown until early stationary phase (4.5-5 hours) and then 5 mL was concentrated into 0.5 mL and incubated with 5 ug pNJ2.1 library DNA in a shaking incubator (250 rpm, 37 °C) for 1 hour. Up to 10 separate cultures were used and pooled during recovery to yield the RS library of >50× coverage. Transformants were selected overnight in LB+chloramphenicol (5ug/mL) to yield the final *B. subtilis* RS library culture.

P. aeruginosa PAO1 was transformed by washing 10 mL of a library overnight culture twice with 300 mM sucrose at room temperature and performing the same final suspension, electroporation, and recovery as with *E. coli* MG1655. A single 1:50 selection was performed in LB Lennox+150 μg/mL carbenicillin at 30 °C, 250 rpm, while taking care not to overgrow the culture and induce biofilm formation or stress responses. Glycerol stocks of all library cultures in final host strains were made upon reaching stationary phase after liquid selection. These stocks were used for all subsequent experiments.

For RS241 library experiments, S. enterica was transformed using the same protocol used for *E. coli*. *V. natriegens* and *C. glutamicum* were transformed according to previously published work[36,47].

Library Growth, DNA-seq, and RNA-seq

For each species, library overnight cultures were made from frozen stocks by diluting 1 mL of thawed frozen stock into 25 mL LB Lennox+antibiotic and grown for 9 hours at 30 °C, 250 rpm. A 1 mL aliquot of this culture was added to 200 mL of pre-warmed LB Lennox and grown (37 °C, 250 rpm) to an OD_{600} of 0.3-0.4 and immediately cooled in an ice slurry. Four 50-mL aliquots were pelleted at 4 °C and the supernatant was removed. Two pellets were resuspended in 5 mL RNAprotect (Qiagen), incubated for 5 minutes at room temperature and repelleted prior to RNA isolation. An additional cell pellet was used for plasmid DNA extraction using a MidiPrep kit (Qiagen) or genomic DNA extraction (only *B. subtilis*, Epicentre MasterPure Gram Positive DNA Purification Kit).

Total RNA was extracted using a Qiagen RNeasy Midi Kit for *E. coli* and *P. aeruginosa* and a modified chemical genomic DNA extraction kit (Epicentre) where the RNase digestion step was replaced with DNase digestion for *B. subtilis*. For *E. coli* alternative growth condition experiments (iron starvation, osmotic stress, minimal media), overnight cultures of the *E. coli* library were pelleted, washed once with PBS, and 1 mL was diluted into 200 mL of LB+200 uM 2,2 dipyridyl (Sigma-Aldrich), LB+0.3 M NaCl, and M9 + glucose. For each condition, pellets were frozen from cultures at OD_{600} 0.3 except for stationary phase library, which was removed at OD_{600} 2.

For RNA-seq library preparation, ribosomal RNA was removed from 4.5 μg of total RNA using Ribo-Zero rRNA Magnetic Removal Kits for gram-negative and gram-positive bacteria (Epicentre). The isolated mRNA was then dephosphorylated using 5′ RNA Polyphosphatase (Epicentre) as follows:

12 uL RNA from previous step

2 uL 10× RNA 5′ Polyphosphatase Reaction Buffer

0.5 uL RiboGuard RNase Inhibitor

1 uL RNA 5′ Polyphosphatase (20 units)

4.5 uL RNase-free water

37 °C for 30 minutes

The reaction was then purified using a Qiagen RNeasy MinElute Kit. We then ligated a 5′ oligo (RNA_adaptor) to the monophosphorylated mRNA as follows:

14 uL RNA from previous step

2 uL 250 uM RNA adaptor

2.5 uL 10× Ligase Buffer

2 uL Epicentre T4 RNA Ligase (10 units)

2 uL 10 mM ATP

1 uL RiboGuard RNase Inhibitor

1 uL DMSO

22.5 °C for 3 hours followed by a 10 minute deactivation at 65 °C.

Our RNA Adaptor contains two terminal N bases to reduce ligation bias[48]. Adaptor-ligated RNA was purified using a Qiagen RNeasy MinElute Kit. Selective reverse transcription was performed using an sfGFP primer as follows:

0.2 uL 10 uM RT Primer

12 uL RNA

1 uL 10 mM dNTP mix

65 °C for 5 min then ice for 1 min

The following components were then added to the PCR tube from the last step:

4 uL of 5× First-Strand Buffer (Invitrogen)

1 uL 0.1 M DTT

1 uL RNaseOUT (Invitrogen)

1 uL SuperScript III Reverse Transcriptase (Invitrogen) (200 units)

The reaction was mixed by gentle pipetting and incubated for 1 hr at 55 °C and then inactivated at 70 °C for 15 min.

To create sequencing libraries, either cDNA or plasmid DNA (or genomic DNA for *B. subtilis*) were amplified in a two-step PCR process using NEBNext High-Fidelity Master Mix with added SYBR (Life Technologies) to add adaptor sequences and indexes for Illumina sequencing. All primers used in this study are listed in Suppl. Materials. Amplification one used an equimolar mixture of four reverse primers (sfGFP_reverse_N3-N6) and vector-specific forward primers to obtain even base distributions during read one of

sequencing. PCR reactions were cycled using a CFX96 Touch Real-Time PCR machine (Bio-Rad) until exponential amplification ceased. A second set of 6-8 qPCR cycles added indexes and Illumina P5 and P7 adaptors for paired-end sequencing. Samples were sequenced on Illumina HiSeq and NextSeq platforms using 300 cycle reads (Paired-End). To validate the transcriptional activity of isolate strains, we performed qPCR on total cDNA extracted from mid-log phase cultures using primers specific to sfGFP and the reference gene ihfB using Kapa SYBR Fast qPCR master mix.

FACS-seq experiments

Two staggered library cultures were grown 1 hour apart following the same protocol for growth used for transcriptional analysis described in the previous section. A 50-mL aliquot was pelleted at 4 °C, resuspended in 5 mL ice-cold 5 PBS. Library cultures were sorted using a FACS Aria 2 (BD Biosciences) into 8 log-spaced bins based on GFP fluorescence (FITC-A) using two consecutive sorts into 4 non-adjacent bins. Samples were kept at 4 °C while sorting. The lowest bin corresponded to the range of fluorescence of a no-sfGFP negative control strain prior to sorting. For the first sort, cells were sorted into bins 1, 3, 5, and 7 until bin 1 (lowest) had ~5 million cells. For the second sort, cells were sorted into the remaining bins at the same rate for the same amount of time to ensure the number of cells sorted into each bin was proportional to the fraction of cells found in each fluorescence range in the original population. Sorted bins were grown in 10 mL LB+antibiotic overnight at 30 °C. We then extracted plasmid DNA or genomic DNA from the sorted populations and amplified the RSs using the same two-step process as described in the previous section. Sequencing was performed on Illumina MiSeq, HiSeq and NextSeq platforms. The median fluorescence value of each bin was determined by diluting each of the sorted overnight cultures 1:200 in 3 mL LB Lennox, growing until OD_{600} of 0.3, pelleting, resuspending cells in chilled PBS and measuring sfGFP fluorescence (FITC-A) on a BD Fortessa flow cytometer. These median values were used to calculate protein levels as described in the next sections. Gene expression from isolate strains from each bin were verified for correspondence with FACS-seq measurements by diluting overnight 96 well plate cultures 1:200 and growing until OD_{600} ~0.3, cooling on ice, and then measuring sfGFP fluorescence (FITC-A) using the high throughput attachment of a BD Fortessa flow cytometer.

Processing steps for analysis of next-generation sequencing reads

Using custom python scripts, we first mapped both RNA and DNA reads to designed RS sequences using its unique 12 bp barcode based on the Read 1.1 sequences. We then confirmed this mapping by aligning the Read 2.1 corresponding to each identified RS sequence to its reference sequence using custom R scripts with the Biostrings package. Mismatched Read 1 and Read 2 assignments were removed from the dataset. We expect that the vast majority of removed reads belong to oligo constructs that had errors during library synthesis, which are mainly deletions. We used a scoring matrix to properly align reads to their reference sequencing whereby mismatches, gap openings, gap extensions, and unresolved bases received scores of −3, −3, 10^{-3}, and 10^{-6}, respectively. Perfect DNA reads align starting at position one in the reference and continue until the end of the read. Read 2 for RNA may begin at a variable position as this is indicative of the transcription start site within the construct. For RNA reads, the first two bases of Read 2 were trimmed off to

account for the random bases in our RNA adaptor. After alignment, we filtered out reads containing errors in more than 4 bp from all analysis. Additionally, any RNA reads beginning upstream of the construct (originating from the vector) were filtered out. After all processing we found that 84, 97, and 75% of constructs had at least one read of DNA or RNA in *B. subtilis*, *E. coli*, and *P. aeruginosa* respectively.

Quantifying transcription and translation levels

Relative transcription levels for each construct (T_i) was determined by the abundance RNA and DNA reads originating from each library member, according to the equation:

$$T_i = \frac{R_i / \sum_i R_i}{D_i / \sum_i D_i}$$

R_i and D_i refer to the total number of RNA and DNA reads for a given library member (*i*). To make comparisons across each recipient organism, raw transcriptional values were normalized by the mean value of active (>0 RNA reads) constructs originating from that species included in the library (159 from *B. subtilis*, 231 from *E. coli*, and 268 from *P. aeruginosa*). We excluded constructs containing 0 DNA counts and also those whose RNA and DNA counts summed to less than 15 for most analyses. However, for visualizations of the range of expression of the data we gave constructs with 0 RNA or DNA reads pseudo-values. For Figures 2A, 3A, and 4B data points with 0 DNA reads and >15 RNA reads (135, 373, and 172 constructs for *B. subtilis*, *E. coli* and *P. aeruginosa* respectively) were given pseudo-value for transcription representing the highest value in the range shown, as these are likely constructs that have fitness defects from high expression that have dropped to low abundance in the population. Constructs that were transcriptionally inactive (0 RNA counts, >15 DNA counts) were given a pseudo-value equal to the minimum value in the range shown.

Translation activity calculations are based on established conventions for FACS-seq. In brief, protein levels for each construct were calculated by normalizing each construct's abundance (D_{ij}) in each bin to the number of reads associated with that bin as well as the fraction of cells from the library sorted into it (f_j). This calculation (below) gives us the fractional abundance (a_{ij}) of each construct in each bin:

$$a_{ij} = \frac{f_j \cdot D_{ij} / \sum_i D_{ij}}{\sum_j \left(f_j \cdot D_{ij} / \sum_i D_{ij} \right)}$$

We then use a weighted average to calculate protein levels (P_i) using fractional abundances and the mean fluorescence level of each bin (m_j) obtained by flow cytometry after sorting and regrowth:

$$\log(P_i) = \sum_j a_{ij} \cdot \log(m_j)$$

This calculation is based on log-normal FACS bins consistent with established conventions in the literature[7,9,49]. Lastly, the data was converted to linear scale and normalized to the minimum fluorescence value and multiplied by 10 so that expression could be compared across species.

Transcription start site determination

We identified the transcription start site (TSS) of active constructs by determining the start position of the alignment of read two with the reference sequence for each RNA read. The first two bases were trimmed in order to take account of the two random bases used for efficient adaptor ligation. The fraction of TSS calls that fell within +/− 5 bp of the median value was then determined. To identify instances of multiple TSS we developed an algorithm utilizing the kmeans function in R. Our algorithm starts with a seed of 6 clusters. The number of clusters is reduced by one if two clusters are found within 5bps of each other or if a cluster contains less than 10% of all reads. Cluster centers and number of clusters are returned at convergence.

Determination of 5′end mRNA Structure stability

Free energy of 5′ end RNA structure was compute using *FOLD* function from RNAstructure package[50]. We defined 5′ end from TSS location up to 20 base pairs after translation initiation site. Only promoter classified as single TSS was used for this analysis. Single TSS promoters were defined as promoter om which greater than 80% of RNA reads lie within 5bps of TSS median.

Regulatory motif discovery and analysis

The MEME package[29] was used to identify regulatory motifs in our dataset. The motif presented in this analysis was obtained by selecting sequences that start 50 bp upstream to TSS up to the translation start site. A random set of 200 promoters of the 10% most expressed promoters was selected for motif finding. The FIMO algorithm was used to scan motif PWM and obtain match scores in our library of promoters. A 4^{th} order GC content background was used for both MEME and FIMO steps.

Hierarchical cluster was performed to identify recipient specific motifs. Only promoters with more than 15 total count (sum of RNA and DNA reads) were used for analysis. Expression was rescaled to interval from 0 to 1 in each recipient. The promoters were split in 10 clusters for motif finding. When masking for promoters with sigma70 motif, all promoters with a motif hit in *E. coli* background (motif p-value < 1e-3) were removed from analysis.

Predicting activity from biophysical parameters

We defined a linear regression model that consider sigma70 motif score, promoter GC content as well as 5′end mRNA stability to predict promoter activity. The $-log_{10}(p\text{-}value)$

was used to define motif sigma70 score between promoter and sigma70 binding. For promoters with more than a single motif hit, the maximum value was used as predictor of affinity. Promoters without any hit better than $-log_{10}(motif_{p\text{-}value})>3$ was given a value of 2. Linear regression was predicted using function *lm* from R package *stats*. Only promoters classified as single TSS (over 80% of reads around median TSS), at least a single count for each RNA and DNA reads and a total count (number of RNA plus DNA reads) greater than 15.

Translation efficiency prediction and determination

We predicted the translation efficiency (or the translation initiation strength) of each member of the RS library using the published Ribosomal Binding Site (RBS) calculator Version 1.0 code[34] (https://github.com/hsalis/Ribosome-Binding-Site-Calculator-v1.0). Input sequences for the RBS calculator consisted of the mRNA sequence of each regulatory sequence (RS) starting from the measured TSS position all the way through 50 bp into the GFP sequence (including the unique barcodes). For RSs with multiple measured TSSs, separate mRNA sequences were generated and predicted independently. A predicted total translation efficiency level for each RS was computed by summing all predicted RBS strengths for each of the mRNAs with alternative TSSs. Translation efficiency predictions were done for each recipient species using specified 16S rRNA anti-Shine-Dalgarno sequences (ACCTCCTTA for *E. coli* and *P. aeruginosa;* ACCTCCTTT for *B. subtilis*) on otherwise default parameters of the RBS calculator algorithm. The experimentally determined translation efficiency is calculated by taking the ratio of the measured transcription rate by the GFP protein levels for each RS. Comparison of *in silico* and experimental translation efficiencies was performed on highly transcribed RSs, corresponding to the highest top 15% transcribed sequences (Suppl. Figure S11).

Cross-species genetic circuits (CGC) construction and measurements

Twelve RSs (1-12) were paired together to generate combinations of double bidirectional RS constructs (Figure 5a). Various RS pairs were synthesized and cloned into pNJ6.2 using PstI-HF and transformed into target strains such that mCherry and sfGFP were controlled by separate RSs separated by a terminator. Constructs were Sanger sequenced to check for synthesis errors and validate the correct cloning orientation. In all, 10 cross-species genetic circuit constructs (A-J) were characterized. Overnight cultures of strains harboring these CGCs were diluted 1:200 and grown in a 96-well plate format in a BioTek H1 Synergy plate reader. Fluorescence values for sfGFP (excitation: 485 nm, emission: 528 nm) and mCherry (excitation: 580, emission: 610nm) were normalized by optical density at the time point closest to OD_{600} = 0.3 to determine reporter activity levels.

Statistical methods

Pearson Correlation—Pearson correlation measures the strength and direction of a linear relationship between two variables. The correlation coefficient *r* can range from -1 to 1, with sign indicating positive or negative association and absolute value indicating the strength of the correlation. For example, in Suppl. Figure S3 we used the Pearson correlation to examine

the reproducibility of transcriptional measurements from independent library cultures, which resulted in a r value of 0.88.

Standard deviation—Standard deviation measures the variation of a set of measurements in relation to their mean. Lower values indicate that individual measurements tend to be close to the sample mean. We used standard deviation (displayed as error bars in Suppl. Figure S7) to examine the variability of individual regulatory sequence transcriptional activity levels across five growth conditions.

Standard error of the mean—Standard error of the mean measures how close a sample's mean value is likely to be from the actual population mean. This is done by dividing the standard deviation by the square root of the sample size. This metric was used in Figure 3b (displayed as error bars) to determine the extent to which calculated mean expression values for different sequence feature value windows may deviate from the true mean.

Linear regression—Linear regression model the relationship between the dependent variable transcriptional activity and multiple independent variables representing sequence features (GC content, mRNA secondary structure stability, sigma factor motif strength) as a linear equation. For the results displayed in Figure 3c we used 10% of the expression data as a training set and the remaining 90% as test sets for each species.

Partial correlation—Partial correlation controls the effects of additional parameters when determining the association between two variables. We used partial correlation to determine which parameters were most informative in our linear regression model (Suppl. Figure S10).

Data availability and Accession Codes

The authors declare that the data supporting the findings of this study are available as supplementary data files. Custom code used for data processing is publicly available at the following link: https://github.com/nathanjohns/PromoterMining. Raw sequencing data can be found at NCBI SRA Bioproject 431139.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Wang lab for helpful discussions and feedback. H.H.W. acknowledges funding support from the NIH (1DP5OD009172-02, 1U01GM110714-01A1), NSF (MCB-1453219), Sloan Foundation (FR-2015-65795), DARPA (W911NF-15-2-0065), and ONR (N00014-15-1-2704). N.I.J. is supported by a NSF Graduate Research Fellowship (DGE-16-44869). S.S.Y. thanks support from Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2017R1A6A3A03003401). We also thank T. Seto for help with plasmid construction, A. Figueroa for assistance with cell sorting, H. Salis for helpful discussions regarding the RBS calculator, D.B. Goodman for discussions regarding FACS-seq, as well as D. Dubnau, S. Lory and A. Rasouly for providing the BD3182 and PAO1 Δ*psy2* strains.

References

1. Brophy JA, Voigt CA. Principles of genetic circuit design. Nat Methods. 2014; 11:508–520. DOI: 10.1038/nmeth.2926 [PubMed: 24781324]
2. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. Nat Methods. 2014; 11:499–507. DOI: 10.1038/nmeth.2918 [PubMed: 24781323]
3. Bayer TS, et al. Synthesis of methyl halides from biomass using engineered microbes. J Am Chem Soc. 2009; 131:6508–6515. DOI: 10.1021/ja809461u [PubMed: 19378995]
4. Stanton BC, et al. Genomic mining of prokaryotic repressors for orthogonal logic gates. Nat Chem Biol. 2014; 10:99–105. DOI: 10.1038/nchembio.1411 [PubMed: 24316737]
5. Rhodius VA, et al. Design of orthogonal genetic switches based on a crosstalk map of sigmas, anti-sigmas, and promoters. Mol Syst Biol. 2013; 9:702. [PubMed: 24169405]
6. Kinney JB, Murugan A, Callan CG Jr, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A. 2010; 107:9158–9163. DOI: 10.1073/pnas.1004290107 [PubMed: 20439748]
7. Kosuri S, et al. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc Natl Acad Sci U S A. 2013; 110:14024–14029. DOI: 10.1073/pnas.1301301110 [PubMed: 23924614]
8. Mutalik VK, et al. Quantitative estimation of activity and quality for collections of functional genetic elements. Nat Methods. 2013; 10:347–353. DOI: 10.1038/nmeth.2403 [PubMed: 23474467]
9. Sharon E, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012; 30:521–530. DOI: 10.1038/nbt.2205 [PubMed: 22609971]
10. Alper H, Fischer C, Nevoigt E, Stephanopoulos G. Tuning genetic control through promoter engineering. Proc Natl Acad Sci U S A. 2005; 102:12678–12683. DOI: 10.1073/pnas.0504604102 [PubMed: 16123130]
11. Mutalik VK, et al. Precise and reliable gene expression via standard transcription and translation initiation elements. Nat Methods. 2013; 10:354–360. DOI: 10.1038/nmeth.2404 [PubMed: 23474465]
12. Lutz R, Bujard H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res. 1997; 25:1203–1210. [PubMed: 9092630]
13. Kang MK, et al. Synthetic biology platform of CoryneBrick vectors for gene expression in Corynebacterium glutamicum and its application to xylose utilization. Appl Microbiol Biotechnol. 2014; 98:5991–6002. DOI: 10.1007/s00253-014-5714-7 [PubMed: 24706215]
14. Tauer C, Heinl S, Egger E, Heiss S, Grabherr R. Tuning constitutive recombinant gene expression in Lactobacillus plantarum. Microb Cell Fact. 2014; 13:150. [PubMed: 25410118]
15. Song Y, et al. Promoter Screening from Bacillus subtilis in Various Conditions Hunting for Synthetic Biology and Industrial Applications. PLoS One. 2016; 11:e0158447. [PubMed: 27380260]
16. Markley AL, Begemann MB, Clarke RE, Gordon GC, Pfleger BF. Synthetic biology toolbox for controlling gene expression in the cyanobacterium Synechococcus sp. strain PCC 7002. ACS Synth Biol. 2015; 4:595–603. DOI: 10.1021/sb500260k [PubMed: 25216157]
17. Elmore JR, Furches A, Wolff GN, Gorday K, Guss AM. Development of a high efficiency integration system and promoter library for rapid modification of Pseudomonas putida KT2440. Metabolic Engineering Communications. 2017; 5:1–8. DOI: 10.1016/j.meteno.2017.04.001 [PubMed: 29188179]
18. Guiziou S, et al. A part toolbox to tune genetic expression in Bacillus subtilis. Nucleic Acids Res. 2016; 44:7495–7508. DOI: 10.1093/nar/gkw624 [PubMed: 27402159]
19. Cardinale S, Arkin AP. Contextualizing context for synthetic biology–identifying causes of failure of synthetic biological systems. Biotechnol J. 2012; 7:856–866. DOI: 10.1002/biot.201200085 [PubMed: 22649052]

20. Temme K, Hill R, Segall-Shapiro TH, Moser F, Voigt CA. Modular control of multiple pathways using engineered orthogonal T7 polymerases. Nucleic Acids Res. 2012; 40:8773–8781. DOI: 10.1093/nar/gks597 [PubMed: 22743271]
21. Kushwaha M, Salis HM. A portable expression resource for engineering cross-species genetic circuits and pathways. Nat Commun. 2015; 6:7832. [PubMed: 26184393]
22. Gaida SM, et al. Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. Nat Commun. 2015; 6:7045. [PubMed: 25944046]
23. Sheth RU, Cabral V, Chen SP, Wang HH. Manipulating Bacterial Communities by in situ Microbiome Engineering. Trends Genet. 2016; 32:189–200. DOI: 10.1016/j.tig.2016.01.005 [PubMed: 26916078]
24. Kim D, et al. Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. PLoS Genet. 2012; 8:e1002867. [PubMed: 22912590]
25. Boutard M, et al. Global repositioning of transcription start sites in a plant-fermenting bacterium. Nat Commun. 2016; 7:13783. [PubMed: 27982035]
26. Wurtzel O, et al. The single-nucleotide resolution transcriptome of Pseudomonas aeruginosa grown in body temperature. PLoS Pathog. 2012; 8:e1002945. [PubMed: 23028334]
27. Torella JP, et al. Unique nucleotide sequence-guided assembly of repetitive DNA parts for synthetic biology applications. Nat Protoc. 2014; 9:2075–2089. DOI: 10.1038/nprot.2014.145 [PubMed: 25101822]
28. Sleight SC, Bartley BA, Lieviant JA, Sauro HM. Designing and engineering evolutionary robust genetic circuits. J Biol Eng. 2010; 4:12. [PubMed: 21040586]
29. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009; 37:W202–208. DOI: 10.1093/nar/gkp335 [PubMed: 19458158]
30. Ishihama A. Functional modulation of Escherichia coli RNA polymerase. Annual review of microbiology. 2000; 54:499–518. DOI: 10.1146/annurev.micro.54.1.499
31. Browning DF, Busby SJ. The regulation of bacterial transcription initiation. Nature reviews Microbiology. 2004; 2:57–65. DOI: 10.1038/nrmicro787 [PubMed: 15035009]
32. Deutscher MP. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. Nucleic acids research. 2006; 34:659–666. DOI: 10.1093/nar/gkj472 [PubMed: 16452296]
33. Caron MP, et al. Dual-acting riboswitch control of translation initiation and mRNA decay. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:E3444–3453. DOI: 10.1073/pnas.1214024109 [PubMed: 23169642]
34. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol. 2009; 27:946–950. DOI: 10.1038/nbt.1568 [PubMed: 19801975]
35. Kong W, Brovold M, Koeneman BA, Clark-Curtiss J, Curtiss R 3rd. Turning self-destructing Salmonella into a universal DNA vaccine delivery platform. Proc Natl Acad Sci U S A. 2012; 109:19414–19419. DOI: 10.1073/pnas.1217554109 [PubMed: 23129620]
36. Weinstock MT, Hesek ED, Wilson CM, Gibson DG. Vibrio natriegens as a fast-growing host for molecular biology. Nat Methods. 2016; 13:849–851. DOI: 10.1038/nmeth.3970 [PubMed: 27571549]
37. Kalinowski J, et al. The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J Biotechnol. 2003; 104:5–25. [PubMed: 12948626]
38. Bikard D, et al. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol. 2014; 32:1146–1150. DOI: 10.1038/nbt.3043 [PubMed: 25282355]
39. Citorik RJ, Mimee M, Lu TK. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol. 2014; 32:1141–1145. DOI: 10.1038/nbt.3011 [PubMed: 25240928]
40. Gomaa AA, et al. Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. MBio. 2014; 5:e00928–00913. DOI: 10.1128/mBio.00928-13 [PubMed: 24473129]

41. Kotula JW, et al. Programmable bacteria detect and record an environmental signal in the mammalian gut. Proc Natl Acad Sci U S A. 2014; 111:4838–4843. DOI: 10.1073/pnas.1321321111 [PubMed: 24639514]

Online Methods References

42. Guerout-Fleury AM, Frandsen N, Stragier P. Plasmids for ectopic integration in Bacillus subtilis. Gene. 1996; 180:57–61. [PubMed: 8973347]

43. Newman JR, Fuqua C. Broad-host-range expression vectors that carry the L-arabinose-inducible Escherichia coli araBAD promoter and the araC regulator. Gene. 1999; 227:197–203. [PubMed: 10023058]

44. Pedelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS. Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol. 2006; 24:79–88. DOI: 10.1038/nbt1172 [PubMed: 16369541]

45. Markowitz VM, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012; 40:D115–122. DOI: 10.1093/nar/gkr1044 [PubMed: 22194640]

46. LeProust EM, et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res. 2010; 38:2522–2540. DOI: 10.1093/nar/gkq163 [PubMed: 20308161]

47. van der Rest ME, Lange C, Molenaar D. A heat shock following electroporation induces highly efficient transformation of Corynebacterium glutamicum with xenogeneic plasmid DNA. Appl Microbiol Biotechnol. 1999; 52:541–545. [PubMed: 10570802]

48. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011; 39:e141. [PubMed: 21890899]

49. Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. Science. 2013; 342:475–479. DOI: 10.1126/science.1241934 [PubMed: 24072823]

50. Mathews DH. RNA Secondary Structure Analysis Using RNAstructure. Curr Protoc Bioinformatics. 2014; 46:12 16 11–25. DOI: 10.1002/0471250953.bi1206s46

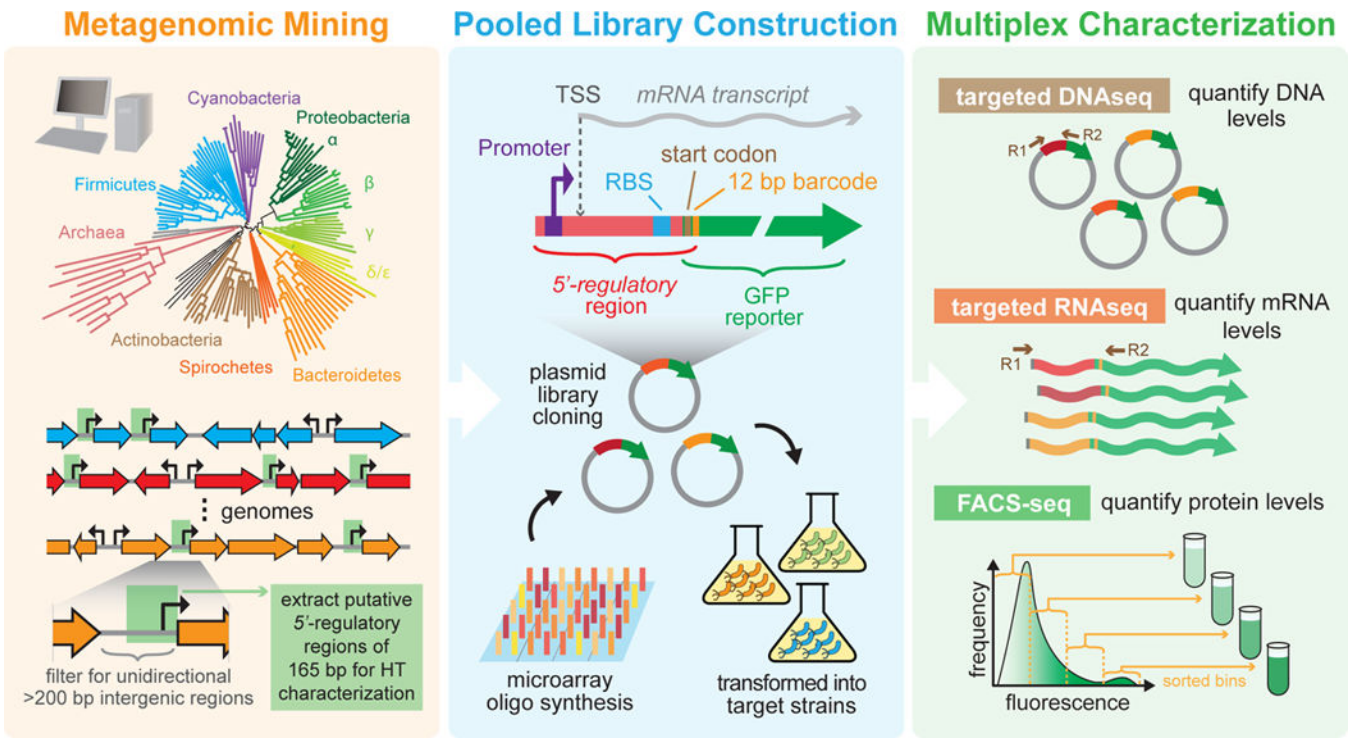

Figure 1.

Metagenomic mining and high-throughput characterization of regulatory sequences from 184 prokaryotic genomes. Unidirectional intergenic regions (>200 bp) were extracted from annotated genomes, trimmed to 165 bp, and assigned unique barcodes, flanking restriction sites, and amplification sequences. The regulatory library was then synthesized on an oligo microarray, amplified, cloned as a pool into species-specific vectors, and transformed into *B.* subtilis, *E. coli*, and *P. aeruginosa* recipients. Targeted RNA-seq, DNA-seq, and FACS-seq enables accurate multiplexed measurement of transcription and translation levels.

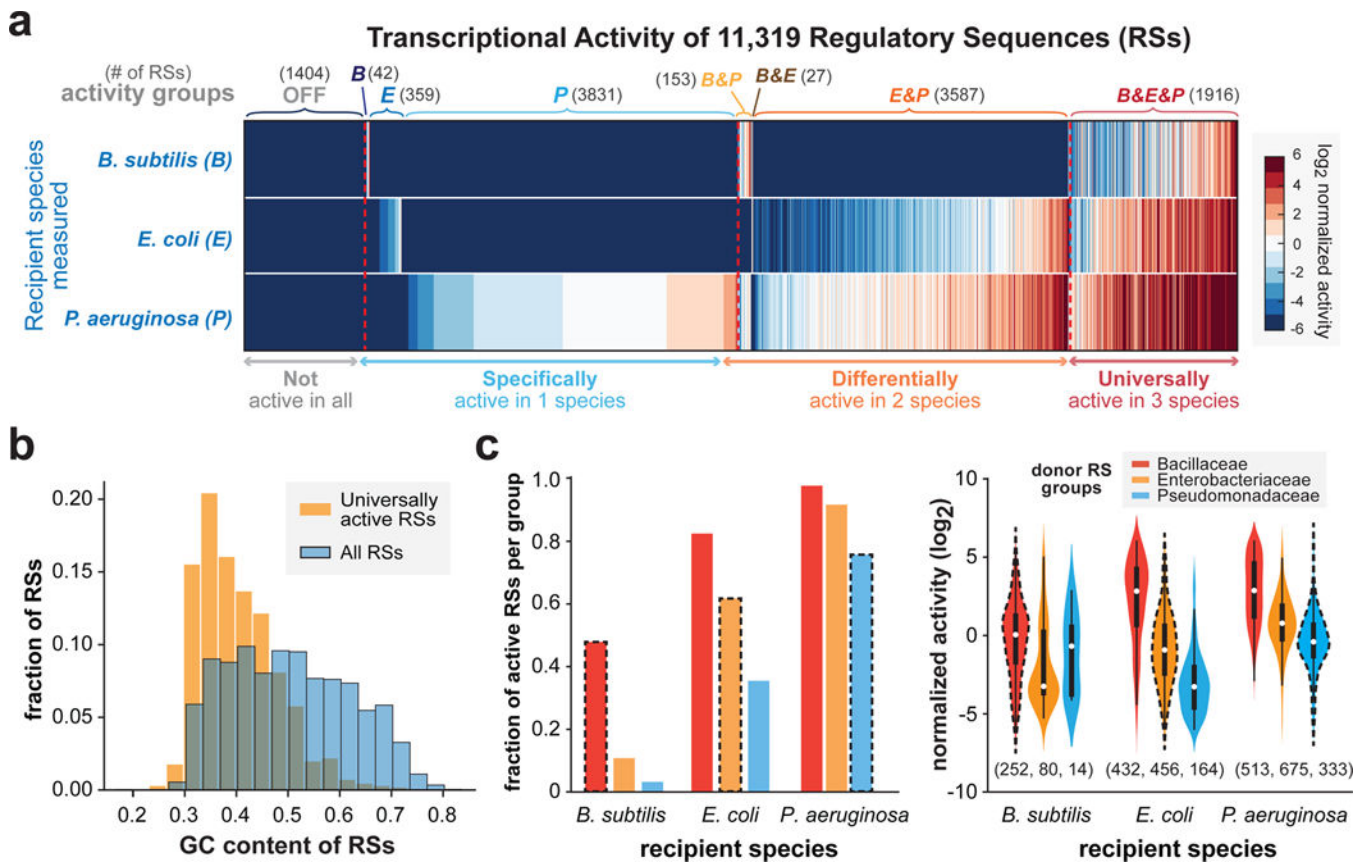

Figure 2.

Transcriptional activities of the regulatory library across 3 diverse species. **(a)** Transcriptional activity of 11,319 regulatory sequences measured in *B.* subtilis, *E. coli*, and *P. aeruginosa* are shown in the heatmap with host-specific groupings annotated above and general categories below. Transcription levels are $\log_2$ (RNA/DNA) ratios normalized by the mean activity of control sequences (see Methods). **(b)** A histogram showing the GC content of the RS library and only the universally active subset, highlighting AT-bias of active RSs. **(c)** The activity profiles of RSs from three distinct phylogenetic groups (red: *Bacillaceae*, orange: *Enterobacteriaceae*, and blue: *Pseudomonadaceae*) measured in each recipient species are shown as fraction active (left) and normalized activity level displayed as a violin plot (right). Box plots (black) with mean values (white dots) are displayed over each violin plot. Cases where donor RS and recipients share the same phylogeny are highlighted in dashed black borders. Sample sizes (n) are listed in parentheses below distributions.

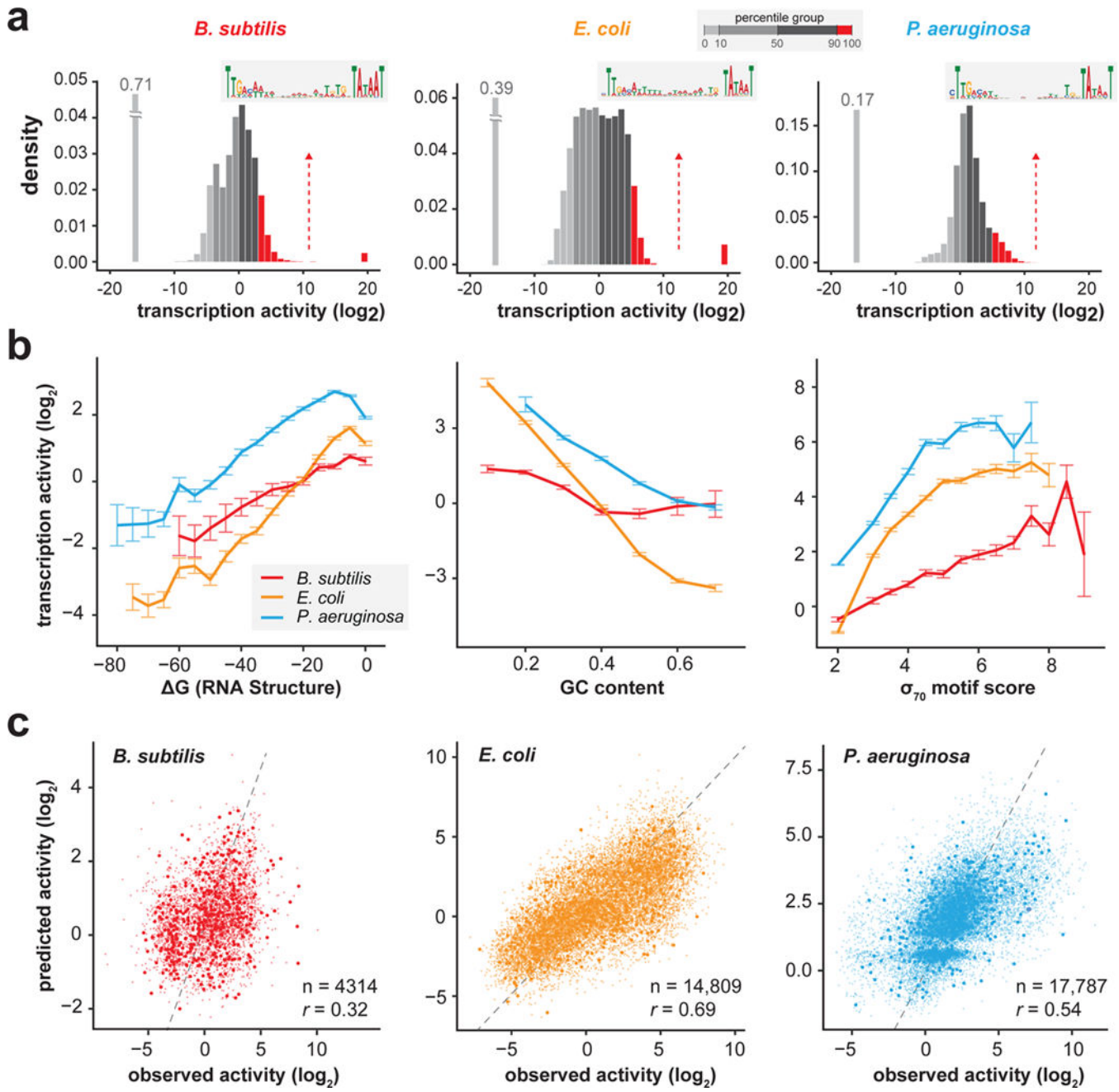

Figure 3.

Assessing regulatory features that govern transcriptional activity. **(a)** Distributions of transcriptional activity is shown for each host. A subset of 200 sequences from the top 10% most active promoters in each recipient were used for separate motif analyses, yielding the dominant *σ70* motif. **(b)** Transcription activity is correlated with biophysical parameters: promoter GC content (left), maximum *σ70* match score (center), mRNA structural stability (right). Mean activities for each feature window are shown with error bars denoting standard errors. **(c)** Linear regression model using the three biophysical parameters. Excluding promoters used to identify the *σ70* motif, the training and test set for the regression model

corresponds to 10% and 90% of the data, respectively. A subset of 500 points is displayed with higher point size to improve visualization. Sample sizes (n) and Pearson correlation coefficients (*r*) are listed in each subplot.

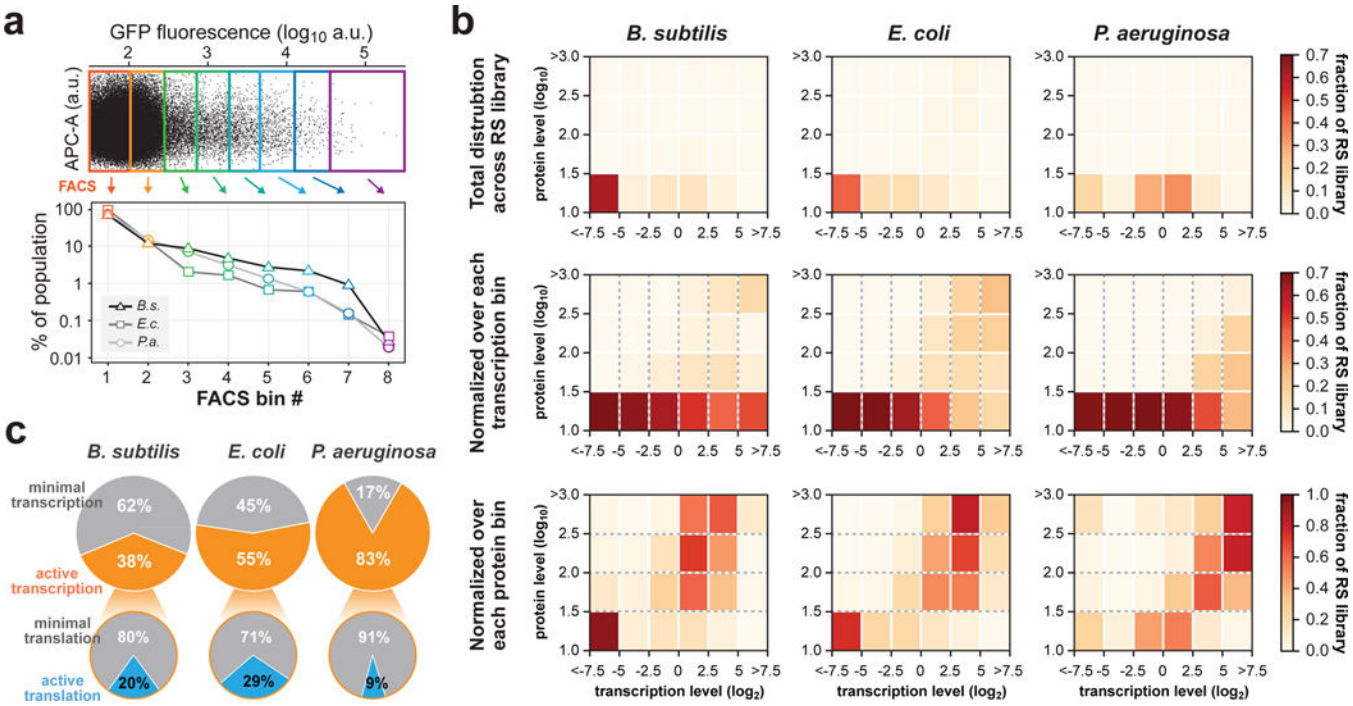

Figure 4.

FACS-seq of RS library. **(a)** Sorting of RS library (top) and the fraction of population sorted into each bin for each host (bottom). **(b)** Heatmap panels show the fraction of RS library distributed across bins of transcription and translation levels in three recipients. The top row of each heatmap subpanels uses values normalized by the total number of regulatory sequences. The middle row uses values normalized by each column bin corresponding to transcription windows. The bottom row uses values normalized by each row bin corresponding to translation windows. **(c)** Pie charts showing fraction of RS library that are transcriptionally active (in orange) and with translational level >1.5 (in blue) based on bins in (b).

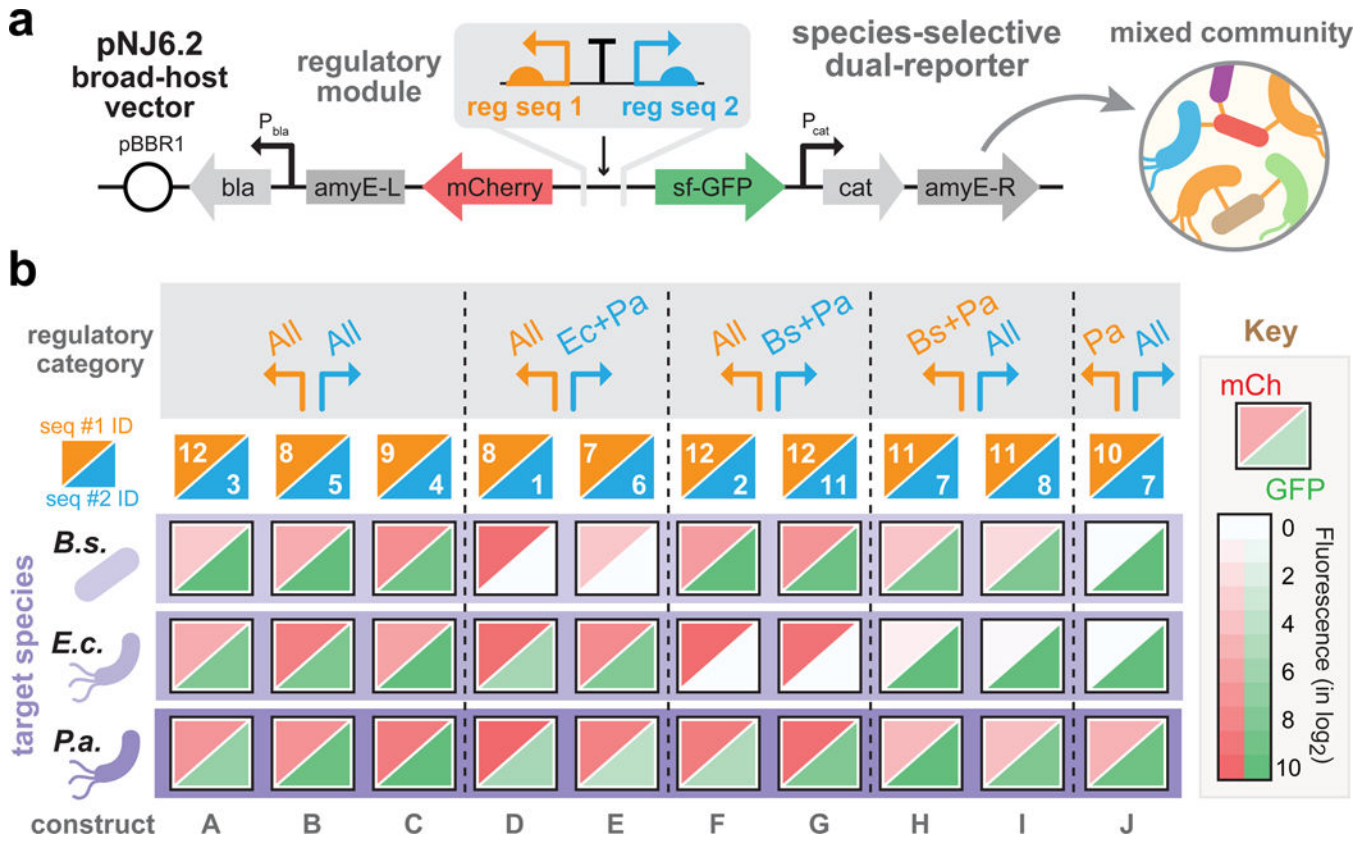

Figure 5.

Species-selective Gene Circuits **(a)** Design of Species-selective Gene Circuits (SsGC) with specified host expression profiles using two outward facing regulatory sequences buffered by a strong bidirectional terminator to drive expression of two fluorescence genes, mCherry and sf-GFP. The pNJ6.2 vector is transformable into *B. subtilis*, *E. coli*, and *P. aeruginosa*. **(b)** Combinatorial construction and fluorescence characterization of 12 host-specified regulatory sequences (Seq ID 1-12) into 10 SsGCs of different regulatory profiles in three recipient species are shown. Distinct regulatory categories include universally active (constructs A-C), *B. subtilis*-excluding or *E. coli*-excluding in the GFP channel (constructs D-E or F-G, respectively), *E. coli*-excluding in the mCherry channel (constructs H-I), and *P. aeruginosa*-specific in the mCherry channel (construct J).