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

Nathan I Johns[1,2,13], Antonio L C Gomes[1,12,13], Sung Sun Yim[1], Anthony Yang[3], Tomasz Blazejewski[1,2], Christopher S Smillie[4], Mark B Smith[5], Eric J Alm[4–7], Sriram Kosuri[8–10] & Harris H Wang[1,11]

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 sequences with broad and host-specific expression properties that are robust in various growth conditions. We also observed substantial differences between species in terms of their capacity to utilize exogenous regulatory sequences. Finally, we demonstrate programmable species-selective gene expression that produces distinct and diverse output patterns in different microbes. Together, these findings provide a rich resource of characterized natural regulatory sequences and a framework that can be used to engineer synthetic gene circuits with unique and tunable cross-species functionality and properties, and also suggest the prospect of ultimately engineering complex behaviors at the community level.

Synthetic biology relies on well-characterized genetic components for the modular assembly of increasingly sophisticated gene circuits with specified functions[1]. Recent advances in high-throughput DNA sequencing and synthesis have greatly increased the ability to generate new genetic parts[2]. Natural enzymes and regulatory proteins have been systematically screened for new functionality[3–5], and noncoding *cis*-regulatory elements have been characterized to improve understanding of their biophysical parameters[6], parts composability[7], contextual robustness[8], and regulatory logic[9] for use in the construction of 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 remains poorly characterized. For many commercially useful microbes, only a handful of regulatory parts have been rigorously tested, and these often have a limited range of expression[13–18]. Efforts to use exogenous regulatory parts in new hosts often fail because of differences in gene expression machinery[19]. More universally compatible and portable regulatory systems have been proposed that use orthogonal regulators[5,20–22], but 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 the use of new types of synthetic circuits to engineer 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 these sequences across different bacterial species and growth conditions and developed species-selective gene circuits with distinct preprogrammed output patterns in different hosts. This data set 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 characterization of natural regulatory sequences

To expand the phylogenetic breadth of useful promoters and translation initiation signals, we first mined 184 prokaryotic genomes for putative regulatory sequences (**Fig. 1**, Online

[1]Department of Systems Biology, Columbia University Medical Center, New York, New York, USA. [2]Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University Medical Center, New York, New York, USA. [3]School of Engineering and Applied Sciences, Columbia University, New York, New York, USA. [4]Broad Institute, Cambridge, Massachusetts, USA. [5]Department of Biological Engineering, MIT, Cambridge, Massachusetts, USA. [6]Computational and Systems Biology Initiative, MIT, Cambridge, Massachusetts, USA. [7]The Center for Microbiome Informatics and Therapeutics, MIT, Cambridge, Massachusetts, USA. [8]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, USA. [9]UCLA–DOE Institute for Genomics and Proteomics, University of California, Los Angeles, Los Angeles, California, USA. [10]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California, USA. [11]Department of Pathology and Cell Biology, Columbia University Medical Center, New York, New York, USA. [12]Present address: Department of Immunology, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [13]These authors contributed equally to this work. Correspondence should be addressed to H.H.W. (hw2429@columbia.edu).

RECEIVED 20 NOVEMBER 2017; ACCEPTED 25 JANUARY 2018; PUBLISHED ONLINE 19 MARCH 2018; DOI:10.1038/NMETH.4633

Methods). These prokaryotes spanned major phylogenetic groups from diverse habitats and included industrially relevant species (**Supplementary Fig. 1**, **Supplementary 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 activity of each RS in the library, we used a previously described high-throughput GFP reporter system[7] (**Fig. 1**). The RS library was generated by microarray oligo synthesis, amplified, cloned as a pool into shuttle vectors (**Supplementary Fig. 2**) upstream of a super-folding GFP (sfGFP), and subsequently transformed into different species for characterization. To determine transcription levels of the RSs in the 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; **Supplementary Fig. 3**). RT-PCR measurements of individual library members were also highly correlated with the corresponding multiplex measurements (**Supplementary Fig. 4a**). To measure translational activity, we used FACS-seq to quantify sfGFP protein levels generated from each RS[6,7] (**Supplementary Fig. 4b**). Flow cytometry measurements of isolated library members showed high correlation with the population-derived FACS-seq library data (**Supplementary Fig. 4c**). Furthermore, transcription and translation measurements obtained with an alternative reporter, mCherry, correlated well with GFP values (**Supplementary Fig. 5**).

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, whereas *E. coli* and *P. aeruginosa* are Gram-negative Proteobacteria that colonize diverse environments. We obtained transcriptional measurements from mid-exponential-phase cultures and generated a converged set of 11,319 regulatory constructs with high-confidence expression across each species. To enable comparisons of transcription profiles between species, we normalized transcription values in each species with endogenous control sequences present in the library, which we used as references to compare activity levels of RSs with those of sequences representative of each host's native transcriptome (Online Methods).

We observed considerable differences in RS transcription activity between different hosts (**Fig. 2a**, **Supplementary Table 2**). *B. subtilis* had fewest measurably active RSs (18.9%, >-6 in $\log_2$), whereas *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 two of three species (33.3%), active in only one species (37.4%), and inactive in all three species (12.4%). In general, universally active RSs had lower GC contents than the overall library (**Fig. 2b**).

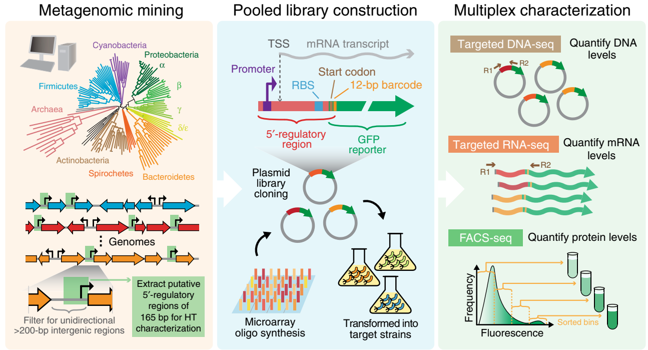

Figure 1 | High-throughput characterization of regulatory sequences from 184 prokaryotic genomes. Unidirectional intergenic regions (>200 bp) were extracted from annotated genomes, trimmed to 165 bp, and assigned unique barcodes, flanking restriction sites, and amplification sequences. The regulatory 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 enable accurate multiplexed measurement of transcription and translation levels.

We observed the converse on the host side, with each organism's capacity to use exogenous RSs appearing to correspond with increasing genomic GC content: *P. aeruginosa* (66% GC) activated the largest fraction of 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 bacterial recipients. We identified distinct patterns of intra- versus inter-family transcriptional specificities (**Fig. 2c**). *B. subtilis* was able to activate 47.7% of donor Bacillaceae RSs, but only 10.8% of Enterobacteracea and 3.2% of Pseudomonaceae RSs. *E. coli* and *P. aeruginosa* were better able to express foreign RSs, 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 (**Fig. 2c**). In contrast, Pseudomonaceae sequences were generally not expressed in *B. subtilis* and were expressed only at low levels in *E. coli*, highlighting the stringent host specificity of its regulatory signals.

We further delineated the regulatory architecture of each sequence by identifying transcription start sites (TSSs) on the basis of our targeted RNA-seq reads. Most TSSs fell between −20 and −50 bp from the start codon (**Supplementary Fig. 6**), in agreement with known native promoter architectures in many bacteria[24–26]. This data set should improve efforts to model bacterial transcription and design new gene circuits. Together, these results highlight that prokaryotic genomes are a rich reservoir of functional regulatory parts with diverse cross-species properties that can be systematically quantified via high-throughput library synthesis and transcriptional profiling.

Because environmental and growth conditions induce changes in gene expression, we also explored the extent to which RS activity is dependent on growth phase or environmental conditions

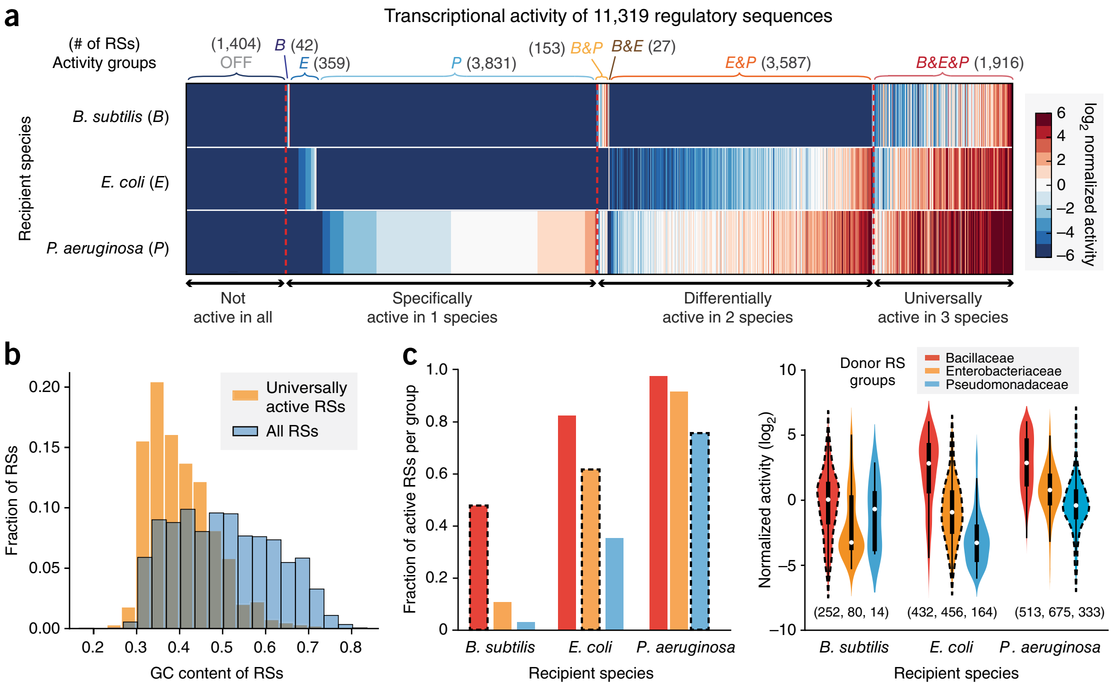

Figure 2 | Transcriptional activity of the regulatory library across three diverse species. (**a**) Transcriptional activity of 11,319 RSs measured in *B. subtilis*, *E. coli*, and *P. aeruginosa*, shown as $\log_2$ (RNA/DNA) ratios normalized by the mean activity of control sequences (Online Methods). Host-specific groupings are annotated above the heat map, and general categories are annotated below it. (**b**) Overlaid histograms of GC content distributions for the RS library and the universally active subset, highlighting the AT bias of active RSs. (**c**) Activity profiles of RSs from three distinct phylogenetic groups measured in each recipient species, shown as the active RS fraction (left) and as normalized activity (right). Box plots (black) displaying the interquartile range (IQR) with median values (white dots) and whiskers extending to the highest and lowest points within 1.5× the IQR are shown over each violin plot. Cases where donor RSs and recipients share the same phylogeny are highlighted by dashed black borders. Sample sizes (*n*) are listed in parentheses below distributions.

experienced by the host. We measured RS transcriptional activity in *E. coli* under five different growth and stress conditions (**Supplementary Fig. 7**, **Supplementary Table 3**). Many RSs (17.3%) showed universally high activity across all conditions, whereas others showed differentially moderate or low transcription activity (28.6% or 22.8%, respectively). TSSs tended to be highly conserved across growth conditions (**Supplementary Figs. 7** and **8**). To generate a set of RSs with robust untranslated regions 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 (**Supplementary Fig. 7d**). We expect this robust RS sublibrary to be a useful resource for designing circuits to be deployed in diverse environments. The use of diverse sequences might 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 identify RS features that govern transcription levels, we carried out *de novo* motif-finding using MEME[29]. For each host, we divided the promoter library into four groups on the basis of activity level (**Supplementary Fig. 9a**). A common motif was enriched in high-activity promoters in all recipients, which corresponded to the canonical binding motif for the housekeeping σ^{70} factor (**Fig. 3a**). Searches for additional motifs yielded only degenerate versions of the core σ^{70} motif (**Supplementary Fig. 9b**,**c**).

To develop a predictive model of transcription activity, we investigated three factors that could influence gene expression: promoter GC content, σ^{70} binding affinity, and 5′ mRNA stability. Promoter GC content indicates compositional preferences of sequence elements that could promote transcription. σ^{70} 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, in combination with the transcription rate, determines overall mRNA transcript levels. Each of the parameters correlated with measured transcription activity of the RS library (**Fig. 3b**). Higher promoter GC content was anticorrelated with transcription activity, whereas a match to the σ^{70} binding motif was positively correlated with activity, as was lower RNA stability (i.e., higher ΔG folding energy). When we controlled for these parameters independently, we determined that the σ^{70} binding motif was most informative for assessments of transcription activity (**Supplementary Fig. 10**). Integration of these parameters into a linear regression model generated predictive powers of 32%, 69%, and 54% for the variances of transcription activity in *B. subtilis*, *E. coli*, and *P. aeruginosa*, respectively (**Fig. 3c**). These results demonstrate that a simple model can explain a considerable fraction of the variation observed in transcriptional activity within different hosts.

Translational activity of regulatory sequences across hosts

Whereas transcriptional activation in bacteria is mediated by transcription factor and σ-factor recruitment of the RNA

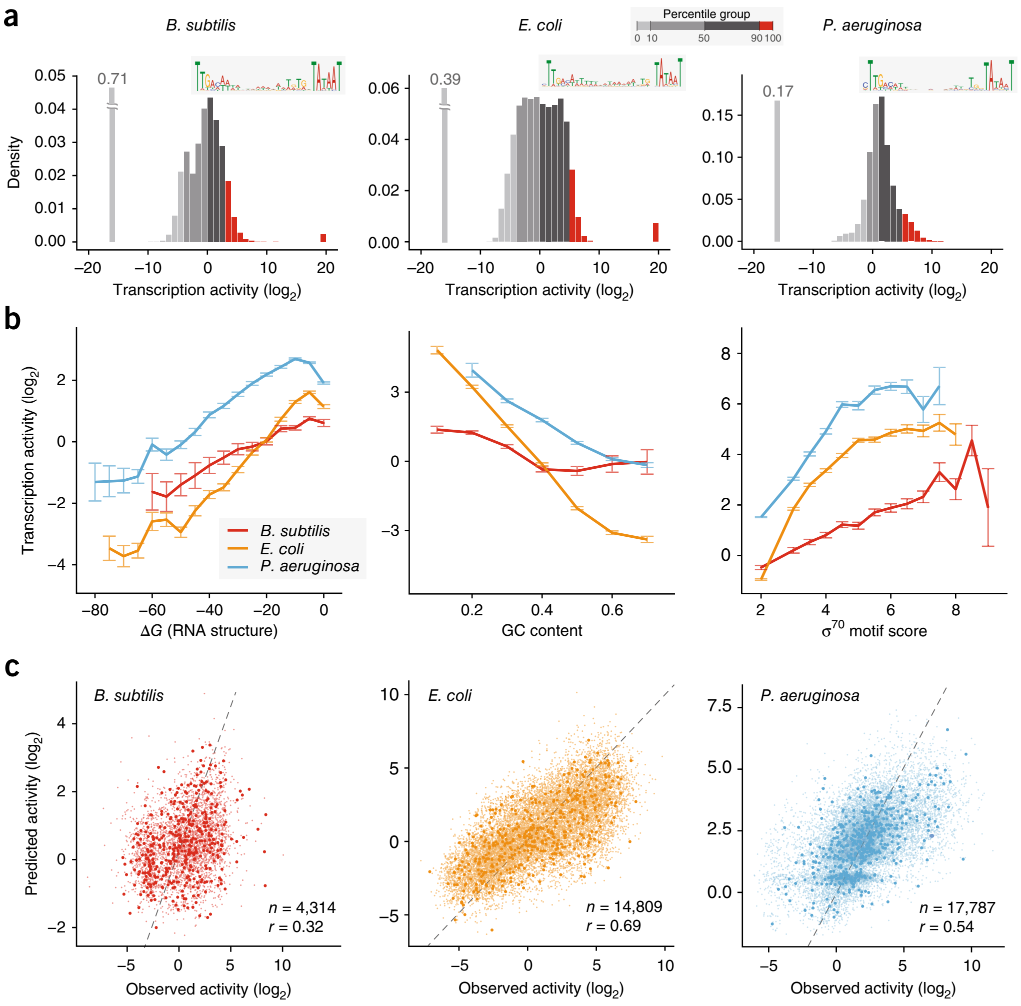

Figure 3 | Assessing regulatory features that govern transcriptional activity. (**a**) Distributions of transcriptional activity for each host. A subset of 200 sequences from the top 10% of the most active promoters (red) in each recipient were used for separate motif analyses, which yielded the dominant σ^{70} motif (sequence logos above plots). (**b**) Transcription activity is correlated with biophysical parameters: mRNA structural stability (left), promoter GC content (center), and maximum σ^{70} match score (right). Activities for each feature window are shown as mean ± s.e., with $n > 10$ for each point shown. Windows with <10 observations are grouped with the nearest neighbor. (**c**) Linear regression model using the three biophysical parameters. Excluding promoters used to identify the σ^{70} motif, the training and test sets for the regression model corresponded to 10% and 90% of the data, respectively. A subset of 500 points is shown with a higher point size for improved visualization. Sample sizes (n) and Pearson correlation coefficients (r) are listed in each plot.

polymerase complex, translational initiation is mediated by interactions between ribosomal subunits and the mRNA transcript. *In silico* modeling of factors that govern ribosomal initiation has allowed 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 used FACS-seq[6,7] to systematically quantify the amount of fluorescence generated from each RS in our library in high throughput across three recipients (**Fig. 4a**). Across the recipients, we identified a shared set of 8,898 RSs that spanned nearly three orders of magnitude of fluorescence (**Supplementary Fig. 11a**), with 3.3% of the library (290 constructs) expressing GFP in all species (**Supplementary Fig. 11b**). Examination of sequences in the region upstream of highly translated library members revealed enrichment of A and G bases centered ~10 bp upstream from the start codon (**Supplementary Fig. 11c**).

To probe the differential effects 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 (**Fig. 4b**). Overall, higher transcriptional activity was associated with higher GFP levels, although translation rates varied widely even for highly transcribed RSs. Normalization over transcription or translation bins highlighted distinct patterns of regulatory specificity associated with RNA or protein generation. RSs belonging to low-transcription bins generally did not yield GFP signal, thus indicating that transcription is a key barrier in gene expression in such cases. Although *P. aeruginosa* was able to transcribe a large fraction of the RS library (83%), only 9% of those RNA species ultimately yielded notable GFP fluorescence, which may reflect incompatibilities at the level of translation (**Fig. 4c**). In contrast, among actively transcribed sequences, *B. subtilis* and *E. coli* were able to yield substantial GFP levels in 20–30% of these RNA

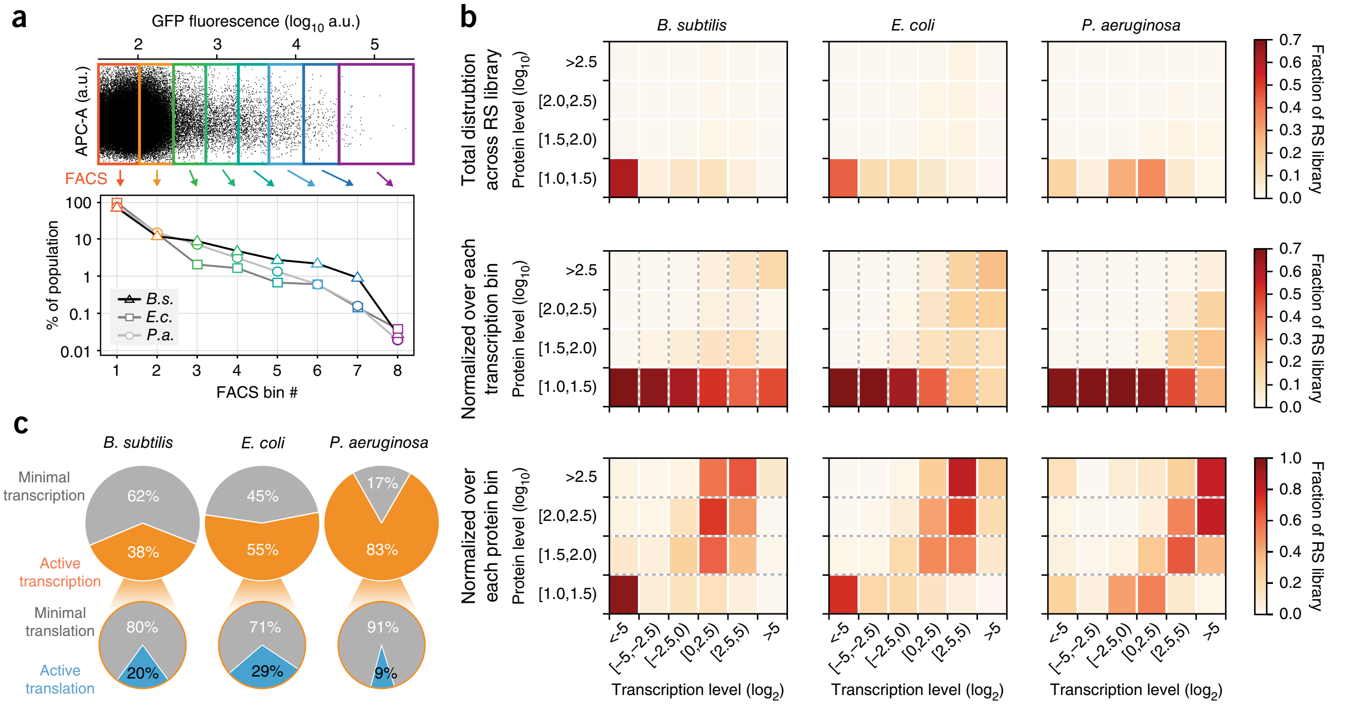

Figure 4 | FACS-seq of RS library. (**a**) Fluorescence distribution and FACS bin organization in the GFP channel versus the allophycocyanin (APC-A) control (top), and the fraction of the population sorted into each bin for each host (bottom). *B.s.*, *B. subtilis*; *E.c.*, *E. coli*; *P.a.*, *P. aeruginosa*. (**b**) Heat maps showing the fraction of the RS library distributed across bins of transcription and translation levels in three bacterial recipients. Values are normalized by the total number of RSs (top row), each column bin corresponding to transcription windows (middle row), or each row bin corresponding to translation windows (bottom row). (**c**) Fractions of the RS library that are transcriptionally active (>0 RNA reads; orange) and have translational levels > 1.5 in log_{10} (blue), based on the bins in **b**, for each host.

transcripts. Interestingly, RSs from Firmicutes species showed high potential to be both transcribed and translated in each host organism (**Supplementary Fig. 12a**). In contrast, although 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 (**Supplementary Fig. 12b**). We additionally assessed the transcription activity and translation efficiency of 212 RSs that contained both active transcription and translation data across all species (**Supplementary Fig. 13a**). We determined the translation efficiency of each RS by normalizing its GFP level to its transcription level. We found that between recipients, only *E. coli* and *P. aeruginosa* showed strong correlations between RSs in terms of transcription levels and translation efficiencies. Finally, we predicted the translation-initiation efficiency of untranslated regions generated from each RS with RBS calculator v1.0 (ref. 34) and found reasonable correlation between predicted values and experimental data (**Supplementary Fig. 13b**).

Together, these results highlight that even if there are similar regulatory specificities at the transcription and translation levels between two species, the two processes have distinct roles in functionalizing heterologous RSs 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 predefined parameters in designs of genetic circuits for use 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 (creating a sublibrary referred to here as RS241), cloned them, and introduced them into the 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 showed activity spanning nearly six orders of magnitude for transcription and three orders of magnitude for translation (**Supplementary Fig. 14**, **Supplementary Table 4**). We observed differential compatibility of RS performance for transcription and translation across phylogenetically diverse species (**Supplementary Fig. 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 of 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 predefined behaviors from the same DNA sequence depending on specificity to the host regulatory machinery (for example, activation of function in only a subset of species). Targeting of subpopulations 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 (SsGCs)

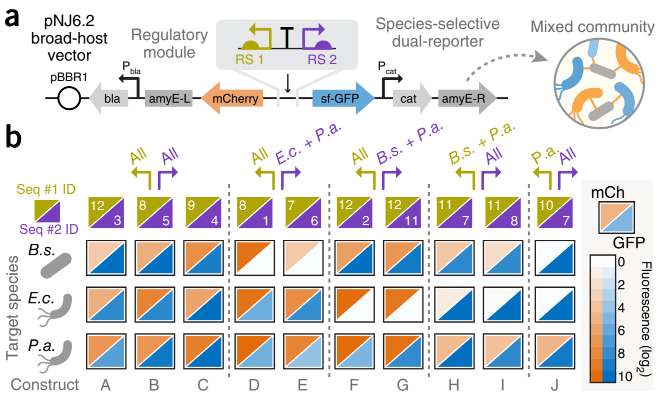

Figure 5 | Species-selective gene circuits. (**a**) Design of SsGCs with specified host expression profiles, using two outward-facing RSs buffered by a strong bidirectional terminator to drive expression of genes encoding two fluorescent proteins, mCherry and sfGFP. The pNJ6.2 vector is transformable into *B. subtilis*, *E. coli*, and *P. aeruginosa*. (**b**) Combinatorial construction and fluorescence characterization of 12 host-specified RSs (Seq IDs 1–12) into ten SsGCs of different regulatory profiles in three recipient species. Distinct regulatory categories include universally active (constructs A–C), *B. subtilis*–excluding or *E. coli*–excluding in the GFP channel (constructs D and E or F and G, respectively), *E. coli*–excluding in the mCherry (mCh) channel (constructs H and I), and *P. aeruginosa*–specific in the mCherry channel (construct J). *B.s.*, *B. subtilis*; *E.c.*, *E. coli*; *P.a.*, *P. aeruginosa*.

that exploit the natural host specificity of heterologous RSs in different bacteria. By leveraging the universal and orthogonal regulatory activation properties observed in our RS library, we built a simple dual-reporter that produced distinct fluorescence states depending on the recipient context (**Fig. 5a**).

We paired 12 RSs to drive a dual mCherry–GFP reporter construct in the broad-host-range vector pNJ6.2, with each regulator independently controlling each fluorescent protein. We introduced each construct into three recipients (*B. subtilis*, *E. coli*, and *P. aeruginosa*) to characterize its host-dependent behaviors. Across ten SsGC constructs (A–J), we demonstrated distinct states of the two reporters, and observed universal, host-specific, and host-excluding activation profiles across recipients (**Fig. 5b**). Some SsGCs exhibited universal activation across all hosts in both reporters (constructs A–C), whereas others had universal activation for mCherry but not sfGFP for *B. subtilis* (constructs D and E). We also built SsGCs that demonstrated the ability to selectively exclude expression of one fluorescent protein in *E. coli* but not in the other species for both reporters (constructs F–I). Additionally, we found that one SsGC induced universal activation of GFP while mCherry expression was limited only to *P. aeruginosa* (construct J), thus 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 toward the generation of more complex functions that could be differentially activated across multiple species of a diverse microbial community, with the ultimate goal of engineering sophisticated community-level dynamics and behaviors.

DISCUSSION

Characterization of regulatory-part performance across different host organisms and growth conditions is crucial for the programming of 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 RSs across up to six diverse bacterial species and under multiple growth conditions. We found substantial differences between species in terms of the ability to transcribe and translate exogenous RSs. 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 transcriptional activation potential—a pattern that appears to be associated with the host species' genomic GC content. We speculate that evolution toward different genomic GC contents may influence the capacity of gene expression machineries to utilize regulatory elements of varying sequence composition. Importantly, we identified and annotated RSs 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 RSs.

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

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

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 an NSF Graduate Research Fellowship (DGE-16-44869). S.S.Y. is supported by the National Research Foundation of Korea (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; G.M. Church (Harvard Medical School, Boston, Massachusetts, USA) for access to OLS libraries; and D. Dubnau

(Rutgers New Jersey Medical School, Newark, New Jersey, USA), S. Lory, and A. Rasouly (both at Harvard Medical School, Boston, Massachusetts, USA) for providing the BD3182 and PAO1 Δ*psy2* strains.

AUTHOR CONTRIBUTIONS

N.I.J., A.L.C.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.C.G., A.Y., T.B., and H.H.W. analyzed the data. N.I.J., A.L.C.G., and H.H.W. wrote the manuscript, with input from all other authors.

COMPETING INTERESTS

The authors declare no competing interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Brophy, J.A. & Voigt, C.A. Principles of genetic circuit design. *Nat. Methods* **11**, 508–520 (2014).
2. Kosuri, S. & Church, G.M. Large-scale *de novo* DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
3. Bayer, T.S. *et al.* Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc.* **131**, 6508–6515 (2009).
4. Stanton, B.C. *et al.* Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat. Chem. Biol.* **10**, 99–105 (2014).
5. Rhodius, V.A. *et al.* Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Mol. Syst. Biol.* **9**, 702 (2013).
6. Kinney, J.B., Murugan, A., Callan, C.G. Jr. & Cox, E.C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. USA* **107**, 9158–9163 (2010).
7. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **110**, 14024–14029 (2013).
8. Mutalik, V.K. *et al.* Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods* **10**, 347–353 (2013).
9. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
10. Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. USA* **102**, 12678–12683 (2005).
11. Mutalik, V.K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
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.* **25**, 1203–1210 (1997).
13. Kang, M.K. *et al.* Synthetic biology platform of CoryneBrick vectors for gene expression in *Corynebacterium glutamicum* and its application to xylose utilization. *Appl. Microbiol. Biotechnol.* **98**, 5991–6002 (2014).
14. Tauer, C., Heinl, S., Egger, E., Heiss, S. & Grabherr, R. Tuning constitutive recombinant gene expression in *Lactobacillus plantarum*. *Microb. Cell Fact.* **13**, 150 (2014).
15. Song, Y. *et al.* Promoter screening from *Bacillus subtilis* in various conditions hunting for synthetic biology and industrial applications. *PLoS One* **11**, e0158447 (2016).
16. Markley, A.L., Begemann, M.B., Clarke, R.E., Gordon, G.C. & Pfleger, B.F. Synthetic biology toolbox for controlling gene expression in the cyanobacterium *Synechococcus* sp. strain PCC 7002. *ACS Synth. Biol.* **4**, 595–603 (2015).
17. Elmore, J.R., Furches, A., Wolff, G.N., Gorday, K. & Guss, A.M. Development of a high efficiency integration system and promoter library for rapid modification of *Pseudomonas putida* KT2440. *Metab. Eng. Commun.* **5**, 1–8 (2017).
18. Guiziou, S. *et al.* A part toolbox to tune genetic expression in *Bacillus subtilis*. *Nucleic Acids Res.* **44**, 7495–7508 (2016).
19. Cardinale, S. & Arkin, A.P. Contextualizing context for synthetic biology—identifying causes of failure of synthetic biological systems. *Biotechnol. J.* **7**, 856–866 (2012).
20. Temme, K., Hill, R., Segall-Shapiro, T.H., Moser, F. & Voigt, C.A. Modular control of multiple pathways using engineered orthogonal T7 polymerases. *Nucleic Acids Res.* **40**, 8773–8781 (2012).
21. Kushwaha, M. & Salis, H.M. A portable expression resource for engineering cross-species genetic circuits and pathways. *Nat. Commun.* **6**, 7832 (2015).
22. Gaida, S.M. *et al.* Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nat. Commun.* **6**, 7045 (2015).
23. Sheth, R.U., Cabral, V., Chen, S.P. & Wang, H.H. Manipulating bacterial communities by in situ microbiome engineering. *Trends Genet.* **32**, 189–200 (2016).
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.* **8**, e1002867 (2012).
25. Boutard, M. *et al.* Global repositioning of transcription start sites in a plant-fermenting bacterium. *Nat. Commun.* **7**, 13783 (2016).
26. Wurtzel, O. *et al.* The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. *PLoS Pathog.* **8**, e1002945 (2012).
27. Torella, J.P. *et al.* Unique nucleotide sequence-guided assembly of repetitive DNA parts for synthetic biology applications. *Nat. Protoc.* **9**, 2075–2089 (2014).
28. Sleight, S.C., Bartley, B.A., Lieviant, J.A. & Sauro, H.M. Designing and engineering evolutionary robust genetic circuits. *J. Biol. Eng.* **4**, 12 (2010).
29. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
30. Ishihama, A. Functional modulation of *Escherichia coli* RNA polymerase. *Annu. Rev. Microbiol.* **54**, 499–518 (2000).
31. Browning, D.F. & Busby, S.J. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2**, 57–65 (2004).
32. Deutscher, M.P. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Res.* **34**, 659–666 (2006).
33. Caron, M.-P. Dual-acting riboswitch control of translation initiation and mRNA decay. *Proc. Natl. Acad. Sci. USA* **109**, E3444–E3453 (2012).
34. Salis, H.M., Mirsky, E.A. & Voigt, C.A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).
35. Kong, W., Brovold, M., Koeneman, B.A., Clark-Curtiss, J. & Curtiss, R. III. Turning self-destructing *Salmonella* into a universal DNA vaccine delivery platform. *Proc. Natl. Acad. Sci. USA* **109**, 19414–19419 (2012).
36. Weinstock, M.T., Hesek, E.D., Wilson, C.M. & Gibson, D.G. *Vibrio natriegens* as a fast-growing host for molecular biology. *Nat. Methods* **13**, 849–851 (2016).
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.* **104**, 5–25 (2003).
38. Bikard, D. *et al.* Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat. Biotechnol.* **32**, 1146–1150 (2014).
39. Citorik, R.J., Mimee, M. & Lu, T.K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nat. Biotechnol.* **32**, 1141–1145 (2014).
40. Gomaa, A.A. *et al.* Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *MBio* **5**, e00928–13 (2014).
41. Kotula, J.W. *et al.* Programmable bacteria detect and record an environmental signal in the mammalian gut. *Proc. Natl. Acad. Sci. USA* **111**, 4838–4843 (2014).

ONLINE METHODS

Bacterial strains and expression vector construction. *E. coli* MegaX DH10B Electrocomp cells (Thermo Fisher; C640003) were used for all initial library cloning steps. Recipient test strains were *E. coli* MG1655, *B. subtilis* BD3182 (a 168 type strain derivative with Δ*rok*::kanR, Met^-, Leu^-, His^- to improve transformation; courtesy of D. Dubnau), and *P. 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, and pNJ3.1 using the backbones pZA11 (p15A ori, 11 copies per cell), pDG1662 (integration into *amyE* locus)[42], and pJN105 (pBBR1 ori, 20 copies per cell)[43], respectively. Unwanted restriction sites for *Pst*I, *Eco*RI, and *Bam*HI found outside of multi-cloning sites were removed by isothermal assembly. An ATG-less sfGFP construct[44] with upstream 5′ *Bam*HI, spacer, *Pst*I, and downstream *Eco*RI was then cloned into each backbone to create the final reporter plasmids (**Supplementary Fig. 3**). We generated the broad-host vector pNJ6.2 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 **Fig. 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 are referred to as RSs for convenience. RSs containing *Bam*HI, *Pst*I, and *Eco*RI sites were filtered out. We randomly chose subsets of RSs from each species, yielding ~160 sequences per genome (**Supplementary Fig. 1**), which led to 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 subselection of the RS sequences. We then added *Bam*HI and *Pst*I 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 effects 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 otherwise stated. The metagenomic RS library was synthesized as a 1-pmol oligo mix by Agilent Technologies (Carlsbad, CA) using their oligo library synthesis platform[46]. The oligo library was first amplified for eight 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 μL of purified amp1 template stock to obtain enough DNA of the library (amp2) for cloning by performing eight 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 *Bam*HI and *Pst*I, 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, and the remaining 990 μL was propagated through two subsequent liquid selections in 25 mL of LB-Lennox (BD Biosciences) + 50 μg/mL carbenicillin grown at 30 °C, 250 r.p.m. overnight. All libraries were cloned with >50× coverage as determined by dividing the number of colony-forming units by the size of the designed library. Plasmid DNA was then extracted from library cultures with 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 of a 100-mL mid-log phase culture with 10% glycerol at 4 °C three times and suspension of the final pellet in 100 μL. Plasmid library DNA (1 μl, 50–100 ng) was added to multiple 20-μL aliquots of competent cells and electroporated at 1.8 kV with a Bio-Rad Micropulser. The cultures were recovered in 1 mL of SOC for 1 h at 30 °C, 250 r.p.m. We determined the library coverage by plating up to 1% of the transformed population on selective plates. The remaining 99% of the transformation culture after 1 h of recovery was passaged through two subsequent liquid selections in 25 mL of LB-Lennox + 50 μg/mL carbenicillin grown at 30 °C, 250 r.p.m. overnight to yield the final *E. coli* RS library.

B. subtilis BD3182 was transformed by dilution of 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 μg/mL of histidine, leucine, and methionine. The culture was grown until early stationary phase (4.5–5 h), and then 5 mL was concentrated into 0.5 mL and incubated with 5 μg of pNJ2.1 library DNA in a shaking incubator (250 r.p.m., 37 °C) for 1 h. Up to ten separate cultures were used and pooled during recovery to yield the RS library of >50× coverage. Transformants were selected overnight in LB + chloramphenicol (5 μg/mL) to yield the final *B. subtilis* RS library culture.

We transformed *P. aeruginosa* PAO1 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 r.p.m., 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 after stationary phase was reached 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].

doi:10.1038/nmeth.4633

Library growth, DNA-seq, and RNA-seq. For each species, we made library overnight cultures from frozen stocks by diluting 1 mL of thawed frozen stock into 25 mL of LB Lennox + antibiotic and growing cultures for 9 h at 30 °C, 250 r.p.m. A 1-mL aliquot of this culture was added to 200 mL of pre-warmed LB Lennox and grown (37 °C, 250 r.p.m.) 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 of RNAprotect (Qiagen), incubated for 5 min at room temperature, and repelleted before 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 with 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 and washed once with PBS, and 1 mL was diluted into 200 mL of LB + 200 µM 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 the 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 with 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 µL of RNA from the previous step
2 µL of 10× RNA 5′ polyphosphatase reaction buffer
0.5 µL of RiboGuard RNase inhibitor
1 µL of RNA 5′ polyphosphatase (20 units)
4.5 µL of RNase-free water
37 °C for 30 min

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

14 µL of RNA from the previous step
2 µL of 250 µM RNA adaptor
2.5 µL of 10× ligase buffer
2 µL of Epicentre T4 RNA ligase (10 units)
2 µL of 10 mM ATP
1 µL of RiboGuard RNase inhibitor
1 µL of DMSO
22.5 °C for 3 h followed by a 10-min deactivation at 65 °C

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

0.2 µL of 10 µM RT primer
12 µL of RNA
1 µL of 10 mM dNTP mix
65 °C for 5 min, then on ice for 1 min

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

4 µL of 5× First-Strand Buffer (Invitrogen)
1 µL of 0.1 M DTT
1 µL of RNaseOUT (Invitrogen)
1 µL of SuperScript III reverse transcriptase (Invitrogen) (200 units)

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

To create sequencing libraries, we amplified either cDNA or plasmid DNA (or genomic DNA for *B. subtilis*) 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 **Supplementary Data Set 1**. Amplification 1 used an equimolar mixture of four reverse primers (sfGFP_reverse_N3-N6) and vector-specific forward primers to obtain even base distributions during read 1 of sequencing. PCR reactions were cycled in 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 h apart according to the same protocol for growth used for transcriptional analysis described in the previous section. A 50-mL aliquot was pelleted at 4 °C, and resuspended in 5 mL of ice-cold 5 PBS. Library cultures were sorted by a FACS Aria 2 (BD Biosciences) into eight log-spaced bins based on GFP fluorescence (FITC-A) using two consecutive sorts into four nonadjacent bins. Samples were kept at 4 °C during sorting. The lowest bin corresponded to the range of fluorescence of a no-sfGFP negative control strain before 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 of 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 done on Illumina MiSeq, HiSeq, and NextSeq platforms. We determined the median fluorescence value of each bin by diluting each of the sorted overnight cultures 1:200 in 3 mL of LB Lennox, growing the culture to an 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. We verified gene expression from isolate strains from each bin for correspondence with FACS-seq measurements by diluting overnight 96-well-plate cultures 1:200, growing them until OD_{600} ~ 0.3, cooling them on ice, and then measuring their 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 RSs using unique 12-bp barcodes based on the Read 1.1 sequences. We then confirmed this mapping by aligning the Read 2.1 corresponding to each identified RS to its reference sequence using custom R scripts with the Biostrings package. Mismatched Read 1 and Read 2 assignments were removed from the data set. We expect that the vast majority of removed reads belonged 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 1 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 TSS 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. The relative transcription level for each construct (T_i) was determined on the basis of the abundance RNA and DNA reads originating from each library member, according to the following equation:

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

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, we normalized raw transcriptional values 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-values for transcription representing the highest value in the range shown, as these are likely to be 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 were based on established conventions for FACS-seq. In brief, we calculated protein levels for each construct by normalizing each construct's abundance ($D_{i,j}$) in each bin to the number of reads associated with that bin and the fraction of cells from the library sorted into it ($f_{,j}$). This calculation (below) gave us the fractional abundance ($a_{i,j}$) of each construct in each bin:

$$a_{i,j} = \frac{(f_j \cdot D_{ij}) / \Sigma_i D_{ij}}{\Sigma_i \left((fi \cdot D_{ij}) / \Sigma_i D_{ij} \right)}$$

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 were 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 TSSs of active constructs by determining the start position of the alignment of read 2 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 TSSs, we developed an algorithm using the kmeans function in R. Our algorithm starts with a seed of six clusters. The number of clusters is reduced by one if two clusters are found within 5 bp of each other or if a cluster contains less than 10% of all reads. Cluster centers and the number of clusters are returned at convergence.

Determination of 5′-end mRNA structure stability. Free energy of 5′-end RNA structure was computed using the FOLD function from the RNAstructure package[50]. We defined the 5′ end from the TSS location up to 20 bp after the translation initiation site. Only promoters classified as single TSS were used for this analysis. Single-TSS promoters were defined as promoters in which >80% of RNA reads lie within 5 bp of the TSS median.

Regulatory motif discovery and analysis. The MEME package[29] was used to identify regulatory motifs in our data set. We obtained the motif presented in this analysis by selecting sequences that start 50 bp upstream of the 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 the motif position weight matrix and obtain match scores in our library of promoters. A fourth-order GC content background was used for both MEME and FIMO steps.

Hierarchical clustering was performed to identify recipient specific motifs. Only promoters with total counts of more than 15 (sum of RNA and DNA reads) were used for analysis. Expression was rescaled to the interval from 0 to 1 in each recipient. The promoters were split into ten clusters for motif finding. When masking for promoters with σ^{70} motifs, all promoters with a motif hit in the *E. coli* background (motif $P < 1 \times 10^{-3}$) were removed from analysis.

Predicting activity from biophysical parameters. We defined a linear regression model that considered σ^{70} motif score, promoter GC content and 5′-end mRNA stability to predict promoter activity. The $-\log_{10}$ of the P value was used to define motif σ^{70} score between promoter and σ^{70} binding. For promoters with more than a single motif hit, the maximum value was used as a predictor of affinity. Promoters without any hit better than $-\log_{10}(P_{\text{motif}}) > 3$ were given a value of 2. Linear regression was predicted using

doi:10.1038/nmeth.4633

the function "lm" from the R package "stats." Only promoters classified as single TSS (over 80% of reads around the median TSS) with at least one count for RNA and DNA reads and a total count (RNA + DNA) > 15, were used in training and test sets.

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 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. We computed a predicted total translation efficiency level for each RS 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. We calculated the experimentally determined translation efficiency 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 carried out on highly transcribed RSs, corresponding to the top 15% of transcribed sequences (**Supplementary Fig. 11**).

Construction and measurement of cross-species genetic circuits. Twelve RSs (1–12) were paired together to generate combinations of double bidirectional RS constructs (**Fig. 5a**). Various RS pairs were synthesized and cloned into pNJ6.2 using *Pst*I-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, ten cross-species genetic circuit constructs (A–J) were characterized. Overnight cultures of strains harboring these cross-species genetic circuits were diluted 1:200 and grown in a 96-well plate in a BioTek H1 Synergy plate reader. Fluorescence values for sfGFP (excitation, 485 nm; emission, 528 nm) and mCherry (excitation, 580 nm; emission, 610 nm) 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 the sign indicating positive or negative association and the absolute value indicating the strength of the correlation. For example, in **Supplementary Figure 3** we use the Pearson correlation to examine the reproducibility of transcriptional measurements from independent library cultures, which resulted in an r value of 0.88.

Standard deviation. The s.d. 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 s.d. (error bars in **Supplementary Fig. 7**) to examine the variability of individual RS transcriptional activity levels across five growth conditions.

Standard error of the mean. The s.e.m. measures how close a sample's mean value is likely to be from the actual population mean. This is done by dividing the s.d. by the square root of the sample size. This metric was used in **Figure 3b** (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 models the relationship between the dependent variable transcriptional activity and multiple independent variables representing sequence features (GC content, mRNA secondary structure stability, σ-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 (**Supplementary Fig. 10**).

Life Sciences Reporting Summary. Further information regarding experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. The data supporting the findings of this study are available as **Supplementary Data Set 1**. Custom code used for data processing is publicly available at GitHub (https://github.com/nathanjohns/PromoterMining). Raw sequencing data can be found at NCBI (SRP131663).

42. Guérout-Fleury, A.M., Frandsen, N. & Stragier, P. Plasmids for ectopic integration in *Bacillus subtilis*. *Gene* **180**, 57–61 (1996).
43. Newman, J.R. & Fuqua, C. Broad-host-range expression vectors that carry the L-arabinose-inducible *Escherichia coli* araBAD promoter and the araC regulator. *Gene* **227**, 197–203 (1999).
44. Pédelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).
45. Markowitz, V.M. *et al.* IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, D115–D122 (2012).
46. LeProust, E.M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).
47. van der Rest, M.E., Lange, C. & Molenaar, D. A heat shock following electroporation induces highly efficient transformation of *Corynebacterium glutamicum* with xenogeneic plasmid DNA. *Appl. Microbiol. Biotechnol.* **52**, 541–545 (1999).
48. Jayaprakash, A.D., Jabado, O., Brown, B.D. & Sachidanandam, R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.* **39**, e141 (2011).
49. Goodman, D.B., Church, G.M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479 (2013).
50. Mathews, D.H. RNA secondary structure analysis using RNAstructure. *Curr. Protoc. Bioinformatics* **46**, 12.6.1–12.6.25 (2014).

doi:10.1038/nmeth.4633

nature research

Corresponding author(s): Harris H. Wang

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

No effect size calculations were performed in this study and therefore no sample size calculations were performed. All sample sizes are listed in each figure's legend.

2. Data exclusions

Describe any data exclusions.

Data exclusions were based on the sequencing coverage of individual constructs and the details are described in our Online Methods.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

To assess reproducibility of our high-throughput data, biological replicates (Supplementary Figure S3) and isolate validations (Supplementary Figure S4) were performed.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was performed

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding was performed

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

□ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ A statement indicating how many times each experiment was replicated

□ ☒ The statistical test(s) used and whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

□ ☒ Test values indicating whether an effect is present
Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.

□ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

□ ☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Custom Python and R scripts.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	There are no restrictions on availability of materials.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	There were no human research participants in this study.

Nature Methods: doi:10.1038/nmeth.4633