HAL
open science

Precise and reliable gene expression via standard transcription and translation initiation elements

Vivek K Mutalik, Joao Guimarães, Guillaume Cambray, Colin Lam, Marc Juul Christoffersen, Quynh-Anh Mai, Andrew B Tran, Morgan Paull, Jay D Keasling, Adam P Arkin, et al.

To cite this version:

Vivek K Mutalik, Joao Guimarães, Guillaume Cambray, Colin Lam, Marc Juul Christoffersen, et al.. Precise and reliable gene expression via standard transcription and translation initiation elements. Nature Methods, 2013, 10 (4), pp.354-360. 10.1038/nmeth.2404 . hal-02950419

HAL Id: hal-02950419

https://hal.inrae.fr/hal-02950419

Submitted on 27 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Precise and reliable gene expression via standard transcription and translation initiation elements

Vivek K Mutalik[1–3], Joao C Guimaraes[1,3,4], Guillaume Cambray[1,3], Colin Lam[1,3], Marc Juul Christoffersen[1,3], Quynh-Anh Mai[1,3], Andrew B Tran[1,3], Morgan Paull[1], Jay D Keasling[1–3,5,6], Adam P Arkin[1–3,8] & Drew Endy[1,7,8]

An inability to reliably predict quantitative behaviors for novel combinations of genetic elements limits the rational engineering of biological systems. We developed an expression cassette architecture for genetic elements controlling transcription and translation initiation in *Escherichia coli*: transcription elements encode a common mRNA start, and translation elements use an overlapping genetic motif found in many natural systems. We engineered libraries of constitutive and repressor-regulated promoters along with translation initiation elements following these definitions. We measured activity distributions for each library and selected elements that collectively resulted in expression across a 1,000-fold observed dynamic range. We studied all combinations of curated elements, demonstrating that arbitrary genes are reliably expressed to within twofold relative target expression windows with ~93% reliability. We expect the genetic element definitions validated here can be collectively expanded to create collections of public-domain standard biological parts that support reliable forward engineering of gene expression at genome scales.

© 2013 Nature America, Inc. All rights reserved.

One main goal of synthetic biology is to make the engineering of biology easier[1,2]. DNA synthesis and assembly has progressed to the point where entire metabolic pathways, chromosomes and genomes can now be synthesized and transplanted[3–5]. However, our capacity to rationally design increasingly complicated genetic systems as enabled by improvements in DNA construction methods has not kept pace[2,6]. One of the greatest claimed barriers to efficient and scalable genetic design is the lack of standard parts that can be reused reliably in novel combinations[6,7]. Many examples instead highlight, even within well-studied organisms such as *E. coli*, how seemingly simple genetic functions behave differently in different settings[8,9]. For example, a prokaryotic ribosome-binding site (RBS) element that initiates translation for one coding sequence might not function at all with another coding sequence[10]. If the genetic elements that encode control of central cellular processes such as transcription and translation cannot be reliably reused, then there is little chance that higher-order objects encoded from such basic elements will be reliable in larger-scale systems[6,11].

Standard biological parts could, in theory, enable hierarchical abstraction of biological functions[1,2,12,13]. The behavior of integrated genetic systems could then be represented via simpler models of individual elements and ultimately mapped to underlying genetic sequences whose encoded functions are dependent on a limited number of measurable or calculable intrinsic variables. Such abstraction of function seems necessary to manage biological complexity and to allow the engineering of increasingly sophisticated genetic systems[6,12,14].

We engineered ~500 transcription and translation initiation elements that are compatible within a standardized genetic context, or expression operating unit (EOU), that enables predictable forward engineering of gene expression over a wide dynamic range. We characterized representative parts for each type by testing more than 1,200 part-part combinations to establish and validate functional composition rules while quantifying scores for part activity. From this data we also estimated the 'quality' of each part, a second-order statistic that represents the extent to which the activity of a part varies across changes in context[15]. Our results demonstrate how, when combined with standardized transcription control elements, a more physically complex design for the control of translation initiation creates simply modeled parts enabling reliable forward engineering of gene expression.

RESULTS

Prioritizing part composition puzzles

In related work, we systematically assembled and tested all combinations of frequently used prokaryotic transcription and translation control elements to quantify average part activities and also variation in activities as parts are reused in novel combinations[15]. Here we focus on developing rules for a genetic layout architecture underlying gene expression cassettes that eliminate

[1]BIOFAB International Open Facility Advancing Biotechnology, Emeryville, California, USA. [2]Lawrence Berkeley National Laboratory, Physical Biosciences Division, Berkeley, California, USA. [3]Department of Bioengineering, University of California, Berkeley, Berkeley, California, USA. [4]Department of Informatics, Computer Science and Technology Center, University of Minho, Campus de Gualtar, Braga, Portugal. [5]Department of Chemical & Biomolecular Engineering, University of California, Berkeley, Berkeley, California, USA. [6]Joint BioEnergy Institute, Emeryville, California, USA. [7]Department of Bioengineering, Stanford University, Stanford, California, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to D.E. (endy@stanford.edu) or A.P.A. (aparkin@lbl.gov).

RECEIVED 30 AUGUST 2012; ACCEPTED 14 FEBRUARY 2013; PUBLISHED ONLINE 10 MARCH 2013; DOI:10.1038/NMETH.2404

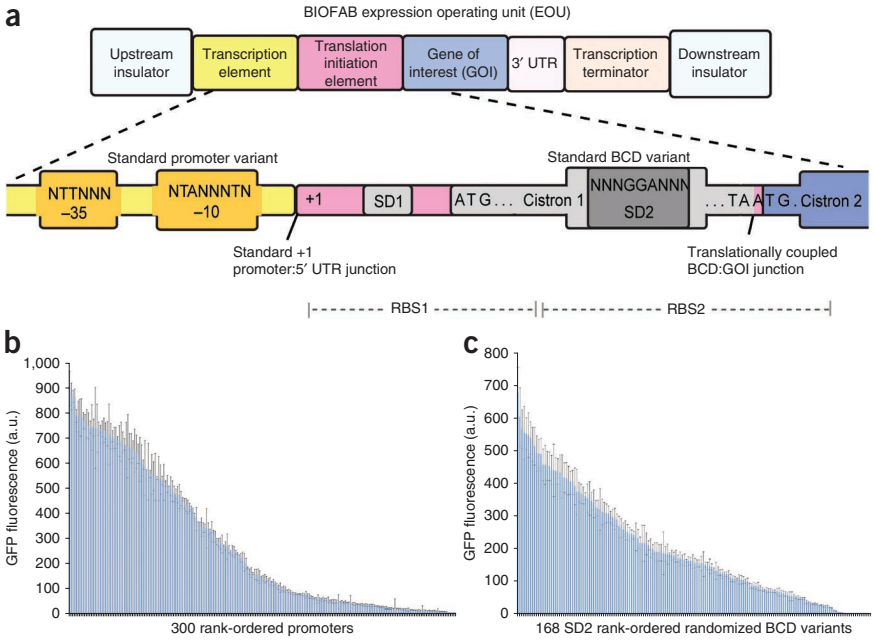

Figure 1 | Rules for regularizing gene expression. (**a**) We defined an expression operating unit (EOU) to set boundaries and junctions of functional genetic elements underlying the expression of heterologous genes (**Supplementary Note**). The variable regions within each element type (wider icons) and the standard junctions (labeled lines) between elements that best enable reliable reuse of elements in novel combinations are detailed. The bicistronic design (BCD) with its two Shine-Dalgarno motifs (SD1 and SD2) is shown. (**b**) Rank-ordered library of constitutive promoters that encode an expected common +1 mRNA boundary and 5′ UTR leader sequence. a.u., arbitrary units. (**c**) Rank-ordered library of SD2 sites that adhere to the BCD and resulting BCD:GOI junction as established here. Error bars, s.d. ($n = 3$).

functional uncertainty arising from the reuse of transcription and translation initiation elements with any gene of interest (GOI) (**Fig. 1**). Although we herein consider only three elements—promoters, 5′ UTRs and GOIs—and two element-element junctions—promoters:5′ UTRs and 5′ UTRs:GOIs (**Fig. 1**)—subsequent work can expand the EOU architecture and variants thereof in a distributed and asynchronous fashion[15].

Recent studies have focused on regularizing a few examples of promoter:5′ UTR junctions via active enzymatic processing of mRNA[16,17]. However, from our prior systematic study of many promoter:5′ UTR and 5′ UTR:GOI combinations, we found that variation in translation initiation rates arising from irregular 5′ UTR:GOI junctions produced most of observed expression irregularities (14% of 17% total)[15]. Given this information and further noting that, in prokaryotes, irregularities arising specifically across 5′ UTR:GOI junctions cannot be eliminated by enzymatic cleavage between a Shine-Dalgarno (SD) sequence and translation start codon, we decided to first pursue the reliable initiation of translation for any gene coding sequence.

Differential formation of mRNA secondary structures spanning 5′ UTR:GOI junctions that then influence ribosome binding or initiation has long been recognized as a major determinant of variation in translation initiation rates[10,18] (**Supplementary Fig. 1**). Given the absence of reliably reusable translation initiation elements, current engineering methods require construction of multiple variant RBSs or recoded coding sequences followed by experimental screening to obtain desired expression levels, presumably through changes in translation initiation efficiency[7,10,19]. For example, the best available computational tool for designing context-optimized translation control elements for use in *E. coli* gives an ~47% chance to design elements that express proteins to within twofold of a target expression level[10]; we note that such quantitative precision in detailing the compositional reliability of designer genetic elements is rare yet necessary to evaluate and improve current engineering practice. However, given current forward-engineering design capacities, if a specific protein expression level is required, then repeated design attempts must be synthesized and tested experimentally, thereby often resulting in combinatorial increases in required design attempts as system complexity increases[2,20].

We instead sought an architecture for 5′ UTR:GOI junctions that would allow an RBS to more reliably encode a distinct and sequence-specific translation initiation rate without sensitivity to variation in the coding sequence of downstream GOIs. We reconsidered past work with difficult-to-express proteins and also reexamined the detailed architecture of natural polycistronic operons[21–27]. Of particular interest were past examples in which a second, independently translated coding sequence is positioned immediately upstream of or slightly overlapping with the coding sequence of any given GOI[22,26]. In such arrangements, the RBS for the GOI is entirely embedded in the coding sequence of the upstream gene, and translation of the downstream cistron might thus be coupled to translation of the upstream cistron[21–26]. More specifically, the intrinsic helicase activity of ribosomes arriving at the stop codon of an upstream cistron might eliminate inhibitory RNA structures that would otherwise disrupt translation initiation of the downstream GOI[21–26,28,29].

To explore whether overlapping genetic elements and active translation coupling might reliably improve translation initiation, we considered genetic designs that encode short leader peptides followed by a downstream GOI[25,26]. One design encodes a 16-amino-acid leader peptide in a first cistron that overlaps by 1 base pair with a variable downstream coding sequence, encoding both a stop and start codon via a −1 frame shift (**Fig. 1a**)[26]. The leader peptide is synthesized by ribosomes that bind to an upstream SD core sequence (SD1); translation of the downstream GOI is thought to result, primarily, from SD1-directed ribosomes that recognize and reinitiate translation via a second SD site (SD2) that is encoded entirely within the coding sequence of the leader peptide[21,22,24,26]. We termed this translational coupling architecture a 'bicistronic design' (BCD) to acknowledge the major difference from conventional 'monocistronic designs' (MCDs), in which translation of coding sequences initiates from an SD site that does not overlap with other functional sequences[25,26]. We found that, unlike SD motifs encoded within MCDs, those encoded within BCDs could initiate protein synthesis even if the

© 2013 Nature America, Inc. All rights reserved.

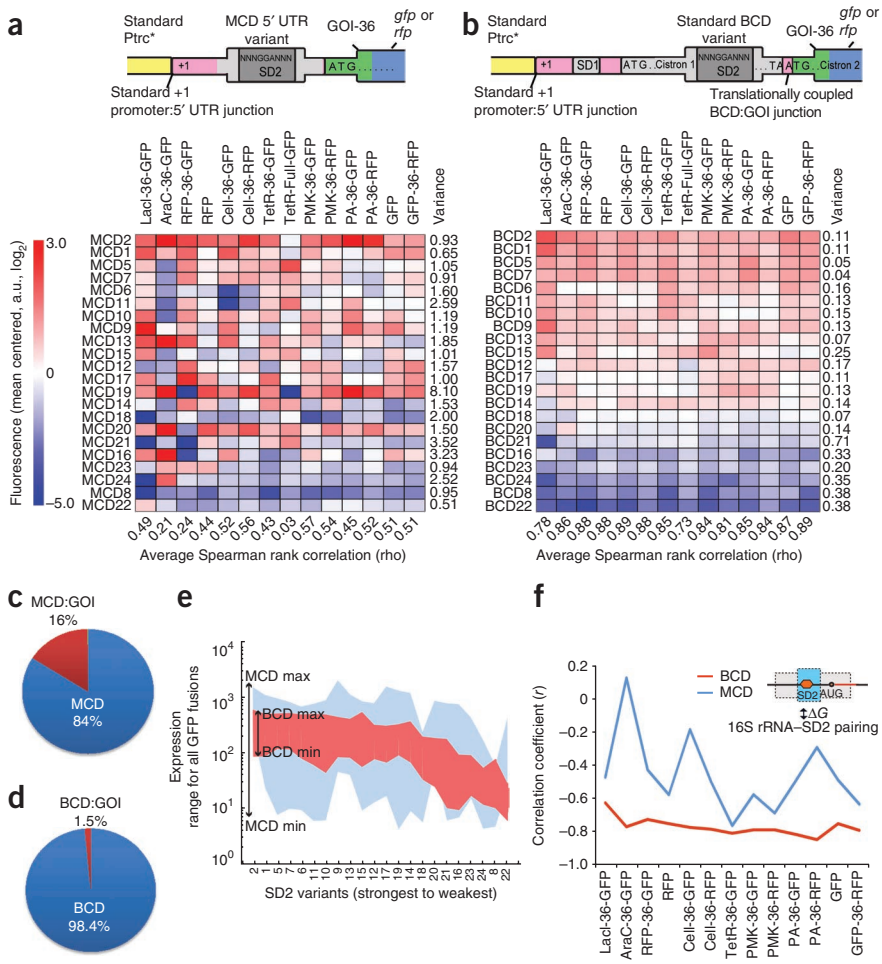

Figure 2 | Standard translation initiation elements using a bicistronic design are reliably reusable. (**a**) Gene expression via a regularized medium-strength promoter (Ptrc; asterisk indicates an absent operator sequence) and 22 monocistronic design (MCD) 5′ UTRs of varying expression strength. Eight GOIs coding for a total of 14 chimeric reporter fusions with either *gfp* or *rfp* (columns) are shown. The 14 chimeric reporter GOIs are encoded via the first 36 nt of the N-terminal coding sequences of *lacI*, *araC*, *rfp*, *gfp*, *tetR* and genes encoding putative cellulase (Cell), phosphomevalonate kinase (PMK) and penicillin acylase (PA) and via the full-length coding sequence of *tetR* (Online Methods). Variance in mean-centered $\log_2$ expression (left) from each MCD across all GOIs sequences (right) and average Spearman rank correlations (bottom) as given (**Supplementary Fig. 8**). a.u., arbitrary units. (**b**) The same SD sequences used in **a** encoded within bicistronic designs (BCDs). Rank orderings for **a** and **b** were established via data of **b**. Variance in mean-centered $\log_2$ expression from each BCD across all GOIs (right) and average Spearman rank correlations (bottom) as given (**Supplementary Fig. 6**). (**c**,**d**) Analysis of variance (Online Methods) in total protein synthesis levels realized using the MCDs (**c**) or BCDs (**d**). (**e**) Comparison of absolute GFP synthesis ranges produced using MCDs or BCDs across all tested GOIs. (**f**) Predicted hybridization free energies between 16S rRNA and SD sequences are better correlated to expression for BCDs than that for MCDs (**Supplementary Figs. 11** and **12**).

coding sequence for the GOI contained a perfect reverse complement to the cognate SD site (**Supplementary Fig. 1**), implying that translation from SD1 disrupts mRNA structure spanning the junction between cistrons such that translation initiation from SD2 is restored.

Precise and reliable translation initiation

We then sought to establish whether the BCD could be generalized so as to initiate synthesis of many proteins across a wide range of translation initiation rates generated by varying the SD sequence to modify differential ribosome-binding affinities[30]. Though the significance of specific SD2 sequence elements has been recognized in a few naturally coupled cistrons[22–24], there are no reports of engineering a library of SD2 variants to fine-tune expression of a downstream GOI. We hypothesized that, for a given SD1 sequence element, a wide range of translation initiation rates could be obtained within a BCD by varying the embedded SD2 sequence. We randomized an SD2 motif, preserving a 3-nucleotide (nt) consensus core, and obtained several hundred sequence-distinct clones encoding a ~600-fold range of reporter-protein expression (**Fig. 1c**, **Supplementary Fig. 2** and Online Methods).

From this BCD library, we chose 22 SD2 candidates of different strengths to test whether each retained its relative encoded strength when used to express sequence-distinct genes (**Supplementary Table 1**). Also, to directly compare the performance of BCDs to conventional MCDs, we used the same SD2 sequences in MCDs. We then assembled a test panel of 14 chimeric reporter GOIs by fusing the first 36 nt (a length thought sufficient to encompass effects of ribosome footprint and mRNA secondary-structure formation on translation initiation[31–33]) from the coding sequences of eight transcription factors or enzymes in-frame to the second codon of a gene encoding GFP or RFP (Online Methods). For added controls, we included a chimeric reporter protein encoded by a full-length *tetR* coding sequence that is fused in-frame to *gfp* in addition to the full-length *gfp* and *rfp* reporter genes (Online Methods). RNA free-energy (ΔG) predictions indicated that our GOI set was expected to form a range of stable mRNA secondary structures spanning BCD:GOI junctions (ΔG from −7 to −24 kcal mol^{-1}; **Supplementary Fig. 3**).

We assembled two full combinatorial test libraries in which 22 MCDs or 22 BCDs were used to translate the 14 chimeric reporter GOIs (Online Methods, **Supplementary Table 1**). We quantified absolute and mean-normalized expression levels by measuring single-cell fluorescence from all 308 MCD:GOI and 308 BCD:GOI combinations (Online Methods, **Fig. 2** and **Supplementary Figs. 4–8**). We observed, as expected, that the synthesis of proteins from conventional MCDs was highly sensitive to changes in the coding sequences of genes (~0.4 average Spearman rank correlation (rho) between any two GOIs; **Fig. 2a** and **Supplementary Figs. 7** and **8**). For example, in a direct comparison of absolute expression, MCD10 driving the *lacI*-36-*gfp* fusion produced ~142-fold more fluorescence than MCD10 driving the *araC*-36-*gfp* fusion, whereas MCD24 driving the

© 2013 Nature America, Inc. All rights reserved.

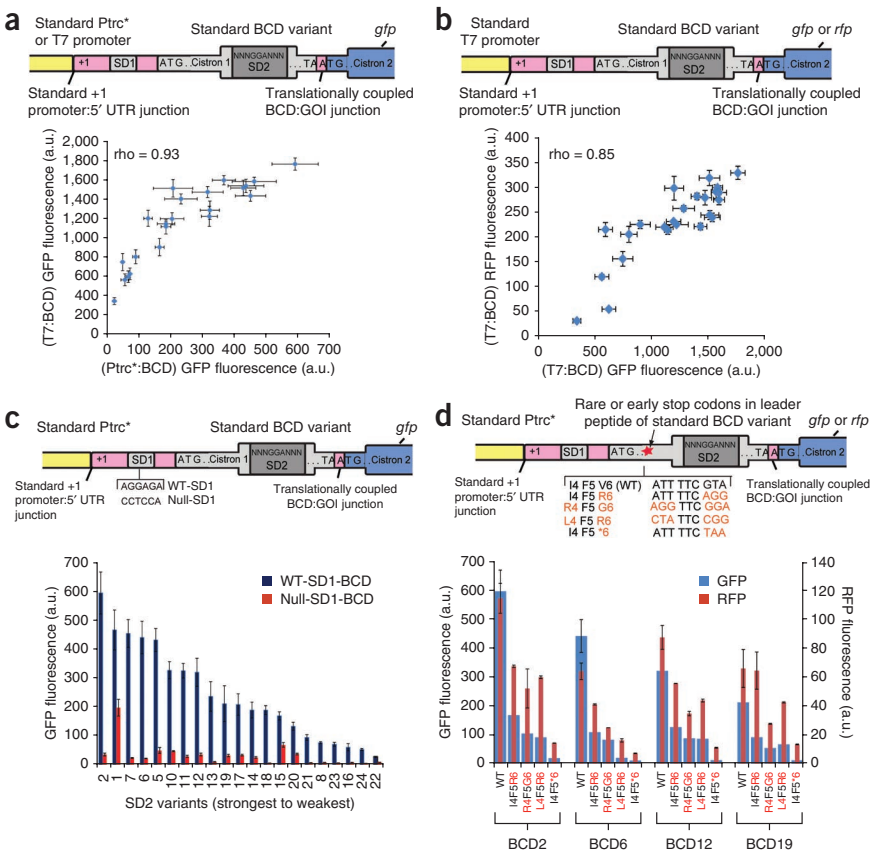

Figure 3 | Bicistronic designs (BCDs) retain functional reliability with alternate transcription systems and different leader cistrons. (**a**) Correlated gene expression levels from BCDs with an *E. coli* Ptrc* promoter (*x* axis) or bacteriophage T7 (*y* axis) promoter and RNA polymerase. The asterisk indicates that the promoter has no operator sequence and hence is constitutive in expression. a.u., arbitrary units. (**b**) Correlated gene expression levels from a phage T7 transcription system but with two GOIs. (**c**) Rank-ordered GFP expression for BCDs (WT-SD1-BCD) compared to expression for those in which SD1 is disrupted (Null-SD1-BCD, schematic). (**d**) Correlated expression levels from an *E. coli* promoter but with stop or rare codons inserted in the BCD leader cistron (schematic) across SD2 elements of different expression strengths (*x* axis, clustered groupings). Error bars, s.d. ($n = 3$).

araC-36-*gfp* fusion produced ~32-fold more fluorescence than MCD24 driving the *lacI*-36-*gfp* fusion (**Supplementary Fig. 7**). In contrast, we observed that the same 22 SD2 motifs, when used within BCDs, maintained their relative fluorescence regardless of the downstream GOI (**Fig. 2b**). For example, BCD10 led to only ~1.5-fold more *lacI*-36-*gfp* than *araC*-36-*gfp* expression, which was achieved by both reducing MCD-mediated *lacI*-36-*gfp* overexpression (~63% decrease) and increasing *araC*-36-*gfp* underexpression (~34-fold increase), as calculated by comparing absolute MCD10- and BCD10-mediated expression levels (**Supplementary Figs. 5** and **7**). Within the BCDs, each SD2 reliably encoded a distinct translation initiation rate across both a wide SD2 activity range and changing GOI contexts (average rho ≈ 0.9 between any two GOIs; **Fig. 2b** and **Supplementary Figs. 5** and **6**). Overall, the BCDs reduced variation in gene expression levels arising from irregularities spanning 5′ UTR:GOI junctions from 16% to 1.5% of the total dynamic range for gene expression (**Fig. 2c**,**d** and **Supplementary Figs. 9** and **10**). These improvements were achieved through the systematic increase of protein synthesis for 5′ UTR:GOI junctions that encoded below-average synthesis levels within an MCD context and the decrease of protein synthesis for 5′ UTR:GOI junctions encoding above-average levels within an MCD context (**Fig. 2e** and **Supplementary Figs. 5** and **7**).

We determined that an equilibrium thermodynamic model based solely on the predicted free energies of binding between 16S rRNA and SD2 sequences is well correlated with observed BCD-mediated protein synthesis (BCD average Pearson correlation coefficient (r) ≈ −0.8 versus MCD average r ≈ −0.4; **Fig. 2f** and **Supplementary Figs. 11** and **12**), further suggesting that the BCD isolates translation initiation activity from variation in downstream gene context. Composite free-energy calculations from a statistical thermodynamic model[10] that considers intermolecular 16S rRNA and mRNA base-pairing as well as other sequence features were less well correlated for BCDs but better correlated for MCDs (average r ≈ −0.6 for both BCD- and MCD-directed protein synthesis; **Supplementary Figs. 13–15**), indicating that the encoded activities of different strength BCDs are best mapped to a relatively simpler core SD2 sequence motif.

We explored whether BCD performance is limited to particular transcription systems or specific internal sequences. First, we used a consensus bacteriophage T7 promoter and polymerase to transcribe BCDs and GOIs. T7 RNA polymerase synthesizes mRNA at a rate up to about eightfold faster than native *E. coli* transcription and translation rates and thus likely results in ribosome-free 5′ mRNA before ribosome loading and translation initiation[34], potentially leading to changes in mRNA folding or processing. We found that the T7 expression system increased average expression levels about fourfold, as expected, and the activities of BCDs remained well correlated to those obtained with a medium strength *E. coli* promoter (rho ≈ 0.9; **Fig. 3a**). We confirmed that the T7 transcription system did not significantly disrupt the reliability of BCDs across changing GOI contexts (rho ≈ 0.9; **Fig. 3b**), whereas the MCDs showed limited reliability in comparison (rho ≈ 0.5; **Supplementary Fig. 16**). We also demonstrated that an active SD1 motif is required to enable reliable initiation at SD2 motifs of different strengths and to translate downstream GOIs (**Fig. 3c**). Such results are in agreement with earlier studies on naturally coupled cistrons in which the significance of varying SD1 has been explored to a limited extent within the context of a stronger and unchanging SD2 sequence[21,24]. We determined that the introduction of rare codons into the leader cistron of the BCD consistently reduced expression levels without major disruptions to the reliable performance of BCDs, and the addition of a stop codon to the leader cistron nearly eliminated expression (**Fig. 3d**, **Supplementary Fig. 17** and Online Methods). Finally, we designed 21 sequence-independent BCDs and confirmed

© 2013 Nature America, Inc. All rights reserved.

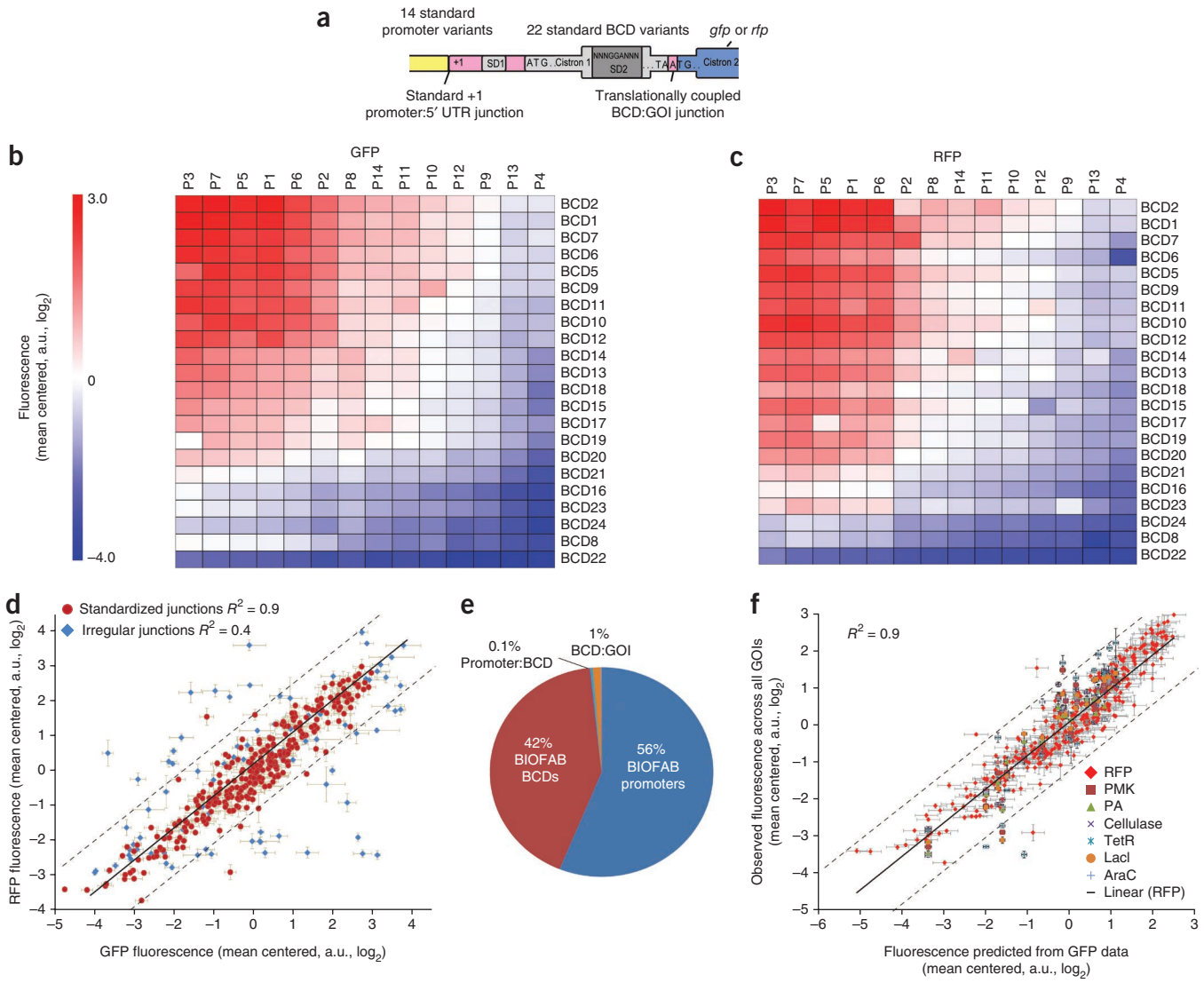

Figure 4 | Precise and reliable gene expression via standard transcription-control and translation-initiation elements. (**a**) Standard promoters produce mRNA from a common +1 nucleotide position. Translation initiation is entirely encoded by a separate and independent bicistronic design (BCD). (**b**,**c**) Mean-centered $\log_2$ expression for green (**b**) and red (**c**) fluorescent proteins via a full combinatorial library of standardized promoters (14) and BCDs (22). a.u., arbitrary units. (**d**) Direct correlation of expression from **b** and **c** (red circles) against those generated by use of irregular transcription- and translation-control elements (blue diamonds, data from ref. 15). (**e**) Factorial analysis of variance for mean-normalized expression from the standard promoter and BCD combinatorial library, with element- and junction-specific contributions to total expression as noted (Online Methods). (**f**) Correlation of observed versus predicted protein expression for sequence-distinct GOIs, as predicted using expression data from a single GOI (GFP) to estimate activity scores for promoters and BCDs adhering to method for forward-engineering gene expression developed here. Error bars: *y* axis, s.d. ($n = 3$); *x* axis, deviations in predicted values derived from the cross-validated model (Online Methods). Cellulase, putative cellulase; PMK, phosphomevalonate kinase; PA, penicillin acylase.

© 2013 Nature America, Inc. All rights reserved.

reliable translation initiation across sequence-distinct GOIs (**Supplementary Figs. 18** and **19** and **Supplementary Table 1**).

Functional composition and reliable gene expression

Building from prior promoter engineering projects[35,36] and transcription initiation studies[37], we chose to regularize promoter:5′ UTR junctions by using promoters that encode a common +1 mRNA start, thereby hoping to avoid complicating requirements such as post-transcriptional mRNA processing[16,17]. We developed a library of variable-strength constitutive promoters with consistent putative mRNA start sites that collectively encoded an ~900-fold dynamic range of expressed reporter levels (**Fig. 1b** and **Supplementary Figs. 20–22**). We selected 14 sequence- and activity-distinct promoters for further study (**Supplementary Table 1** and **Supplementary Figs. 22** and **23**). We assembled each promoter with all 22 BCDs, and we tested expression using two sequence-distinct GOIs (*gfp* and *rfp*; **Fig. 4a** and Online Methods).

We found that the individual rank orderings for promoters and BCDs and resulting GFP or RFP expression levels were systematically maintained and well correlated across a 1,000-fold range for observed protein fluorescence (coefficient of determination (R^2) = 0.9; **Fig. 4b–d** and **Supplementary Figs. 24** and **25**). An analysis of variance of observed fluorescence indicated that

98% of the total dynamic protein expression range was due to encoded differences in the intrinsic activities of individual promoters and BCDs, and not to unknown effects arising from the reuse of these expression control elements in novel combinations (**Fig. 4e**, **Supplementary Fig. 26** and Online Methods). Moreover, a quantitative model for gene expression based only on observed GFP fluorescence levels allowed us to predict observed fluorescence for RFP and other GOIs (R^2 = 0.9; **Fig. 4f** and Online Methods). We also tested the performance of one BCD with variable-strength promoters regulated by one of two popular transcription repressors (**Supplementary Fig. 27** and **Supplementary Table 1**). These results confirmed that BCDs can be used in conjunction with inducible promoters.

DISCUSSION

Users of the genetic elements described above should achieve an ~93% chance to obtain expected GOI-normalized relative expression for a given gene to within twofold of a target level, which represents an ~87% reduction in forward-engineering expression error compared to the error rates of previously best available methods[10] (Online Methods). Our results illustrate that it is possible to overcome many of the challenges thought to limit the engineering of synthetic biological systems via standard biological parts: (i) lack of systematic part characterization, (ii) incompatibility of performance within part collections, (iii) variable part performance across changing genetic contexts and (iv) lack of precise and predictable behavior when used[38]. However, just as one early, reliable screw-thread standard[39] did not itself enable all of mechanical engineering, much work remains in, for example, expanding EOU architectures to incorporate and validate additional genetic functions in *E. coli* and across many organisms.

In establishing reliable promoter:5′ UTR and 5′ UTR:GOI junctions, we used two distinct strategies. The promoter:5′ UTR junction was simply regularized by ensuring that promoters do not contribute mRNA sequence to a standardized 5′ UTR sequence, thereby providing simple functional decoupling. However, rendering a standard and predictably functioning 5′ UTR:GOI junction required a genetic layout in which genetic elements were nested, overlapping and functionally coupled as is common to many natural genetic systems (microbes, phages, viruses and some eukaryotes)[22,23,40,41]. In contrast, designers of early and ongoing synthetic biology 'refactoring' projects have purposefully removed such complexity to enhance physical layout simplicity and presumed functional independence for individual genetic elements[42–44]. We suspect that natural genetic systems might provide further lessons for how more complicated physical couplings can encode simpler and more reliable functional composition schemes.

The BCD could likely be used in combination with other gene expression regulatory elements and designs[45–47] to engineer synthetic polycistronic expression cassettes[48] or to reduce library sizes in directed-evolution efforts by allowing rational choice of a few sequences that cover a desired expression parameter space[49]. Sequence-distinct BCDs are available for engineering multigene systems if genetic instability arising from direct repeats of DNA elements were undesirable (Online Methods, **Supplementary Figs. 18** and **19** and **Supplementary Table 1**). Though we did not observe growth defects or other deleterious phenotypes due to expression of BCD-encoded leader peptides, further studies should consider potential impacts arising from their repeated overexpression. Finally, although research to understand translation initiation in MCD contexts is relatively well established[50], direct observation of how ribosomes reinitiate translation and overcome inhibitory mRNA structures in BCDs, in polycistronic operons and across varying coding sequence contexts would be helpful.

DNA sequence data and functional information detailing the performance of the standard promoters and BCDs established here have been contributed to the public domain and are freely available for use via human- and machine-readable interfaces (http://biofab.org/data/). Potential variation in specific sequence-distinct protein levels due to mechanisms that act downstream of translation initiation must still be accounted for to obtain absolute target protein concentrations[19] (Online Methods). Given an expected 93% reliability rate (7% failure rate) for precision expression engineering, designers of heterologous genetic systems and tool developers working to support the engineering design process[2,6,7,43] might further explore how to best practically enable a priori quantitative specification of desired protein synthesis levels within systems encoding up to about ten genes.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Smolke for discussions. We acknowledge support from a US National Science Foundation grant to the BIOFAB (EEC 0946510) and unrestricted gifts from Genencor, Agilent and DSM. J.C.G. acknowledges financial support from the Portuguese Fundação para a Ciência e a Tecnologia (FCT) (SFRH/BD/47819/2008); G.C. acknowledges the Human Frontier Science Program (LT000873/2011-l) and Bettencourt Schueller Foundation; A.P.A. and D.E. acknowledge the Synthetic Biology Engineering Research Center under National Science Foundation grant 04-570/0540879. This work was conducted at the Joint BioEnergy Institute supported by the Office of Science, Office of Biological and Environmental Research, US Department of Energy, contract DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

V.K.M., A.P.A. and D.E. conceived the study and designed the experiments. V.K.M., C.L., Q.-A.M., A.B.T. and M.P. performed the experiments. V.K.M., J.C.G., G.C., M.J.C., A.P.A. and D.E. analyzed the data. V.K.M., J.C.G., G.C., J.D.K., A.P.A. and D.E. wrote the manuscript. All authors discussed and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
2. Purnick, P.E. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.* **10**, 410–422 (2009).
3. Ellis, T., Adie, T. & Baldwin, G.S. DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol. (Camb.)* **3**, 109–118 (2011).
4. Carr, P.A. & Church, G.M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
5. Gibson, D.G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
6. Lu, T.K., Khalil, A.S. & Collins, J.J. Next-generation synthetic gene networks. *Nat. Biotechnol.* **27**, 1139–1150 (2009).
7. Keasling, J.D. Manufacturing molecules through metabolic engineering. *Science* **330**, 1355–1358 (2010).
8. Cardinale, S. & Arkin, A.P. Contextualizing context for synthetic biology—identifying causes of failure of synthetic biological systems. *Biotechnol. J.* **7**, 856–866 (2012).

© 2013 Nature America, Inc. All rights reserved.

9. Kittleson, J.T., Wu, G.C. & Anderson, J.C. Successes and failures in modular genetic engineering. *Curr. Opin. Chem. Biol.* **16**, 329–336 (2012).
10. 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).
11. Cambray, G., Mutalik, V.K. & Arkin, A.P. Toward rational design of bacterial genomes. *Curr. Opin. Microbiol.* **14**, 624–630 (2011).
12. Canton, B., Labno, A. & Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.* **26**, 787–793 (2008).
13. Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S. & Elowitz, M.B. Accurate prediction of gene feedback circuit behavior from component properties. *Mol. Syst. Biol.* **3**, 143 (2007).
14. Smolke, C.D. Building outside of the box: iGEM and the BioBricks Foundation. *Nat. Biotechnol.* **27**, 1099–1102 (2009).
15. Mutalik, V.K. *et al.* Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods* advance online publication, doi:10.1038/nmeth.2403 (10 March 2013).
16. Lou, C., Stanton, B., Chen, Y.J., Munsky, B. & Voigt, C.A. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.* **30**, 1137–1142 (2012).
17. Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A. & Arkin, A.P. RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.* **30**, 1002–1006 (2012).
18. Dreyfus, M. What constitutes the signal for the initiation of protein synthesis on *Escherichia coli* mRNAs? *J. Mol. Biol.* **204**, 79–94 (1988).
19. Welch, M., Villalobos, A., Gustafsson, C. & Minshull, J. You're one in a googol: optimizing genes for protein expression. *J. R. Soc. Interface* **6** (suppl. 4), S467–S476 (2009).
20. Bonnet, J., Subsoontorn, P. & Endy, D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc. Natl. Acad. Sci. USA* **109**, 8884–8889 (2012).
21. Spanjaard, R.A. & Vanduin, J. Translational reinitiation in the presence and absence of a Shine and Dalgarno sequence. *Nucleic Acids Res.* **17**, 5501–5507 (1989).
22. Oppenheim, D.S. & Yanofsky, C. Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics* **95**, 785–795 (1980).
23. Schümperli, D., McKenney, K., Sobieski, D.A. & Rosenberg, M. Translational coupling at an intercistronic boundary of the *Escherichia coli* galactose operon. *Cell* **30**, 865–871 (1982).
24. Das, A. & Yanofsky, C. A ribosome binding site sequence is necessary for efficient expression of the distal gene of a translationally-coupled gene pair. *Nucleic Acids Res.* **12**, 4757–4768 (1984).
25. Schoner, B.E., Belagaje, R.M. & Schoner, R.G. Translation of a synthetic two-cistron mRNA in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **83**, 8506–8510 (1986).
26. Makoff, A.J. & Smallwood, A.E. The use of two-cistron constructions in improving the expression of a heterologous gene in *E. coli*. *Nucleic Acids Res.* **18**, 1711–1718 (1990).
27. Mendez-Perez, D., Gunasekaran, S., Orler, V.J. & Pfleger, B.F. A translation-coupling DNA cassette for monitoring protein translation in *Escherichia coli*. *Metab. Eng.* **14**, 298–305 (2012).
28. Takyar, S., Hickerson, R.P. & Noller, H.F. mRNA helicase activity of the ribosome. *Cell* **120**, 49–58 (2005).
29. Qu, X. *et al.* The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475**, 118–121 (2011).
30. Barrick, D. *et al.* Quantitative analysis of ribosome binding sites in *E.coli*. *Nucleic Acids Res.* **22**, 1287–1295 (1994).
31. Steitz, J.A. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature* **224**, 957–964 (1969).
32. Yusupova, G.Z., Yusupov, M.M., Cate, J.H. & Noller, H.F. The path of messenger RNA through the ribosome. *Cell* **106**, 233–241 (2001).
33. Kudla, G., Murray, A.W., Tollervey, D. & Plotkin, J.B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
34. Iost, I., Guillerez, J. & Dreyfus, M. Bacteriophage T7 RNA polymerase travels far ahead of ribosomes *in vivo*. *J. Bacteriol.* **174**, 619–622 (1992).
35. Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. USA* **102**, 12678–12683 (2005).
36. Cox, R.S. III, Surette, M.G. & Elowitz, M.B. Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* **3**, 145 (2007).
37. Hook-Barnard, I.G. & Hinton, D.M. Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene Regul. Syst. Bio.* **1**, 275–293 (2007).
38. Kwok, R. Five hard truths for synthetic biology. *Nature* **463**, 288–290 (2010).
39. Sellers, W. A system of screw threads and nuts. *J. Franklin Inst.* **77**, 344–350 (1864).
40. Kozak, M. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**, 187–208 (1999).
41. Scherbakov, D.V. & Garber, M.B. Overlapping genes in bacterial and phage genomes. *Mol. Biol.* **34**, 485–495 (2000).
42. Chan, L.Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol. Syst. Biol.* **1**, 2005.0018 (2005).
43. Temme, K., Zhao, D. & Voigt, C.A. Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl. Acad. Sci. USA* **109**, 7085–7090 (2012).
44. Jaschke, P.R., Lieberman, E.K., Rodriguez, J., Sierra, A. & Endy, D. A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. *Virology* (2012).
45. Mutalik, V.K., Qi, L., Guimaraes, J.C., Lucks, J.B. & Arkin, A.P. Rationally designed families of orthogonal RNA regulators of translation. *Nat. Chem. Biol.* **8**, 447–454, **434**, 278–284 (2012).
46. Liu, C.C., Qi, L., Yanofsky, C. & Arkin, A.P. Regulation of transcription by unnatural amino acids. *Nat. Biotechnol.* **29**, 164–168 (2011).
47. Chang, A.L., Wolf, J.J. & Smolke, C.D. Synthetic RNA switches as a tool for temporal and spatial control over gene expression. *Curr. Opin. Biotechnol.* **23**, 679–688 (2012).
48. Pfleger, B.F., Pitera, D.J., Smolke, C.D. & Keasling, J.D. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.* **24**, 1027–1032 (2006).
49. Cobb, R.E., Si, T. & Zhao, H. Directed evolution: an evolving and enabling synthetic biology tool. *Curr. Opin. Chem. Biol.* **16**, 285–291 (2012).
50. Aitken, C.E., Petrov, A. & Puglisi, J.D. Single ribosome dynamics and the mechanism of translation. *Annu. Rev. Biophys.* **39**, 491–513 (2010).

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

Bacterial strains, plasmids and growth conditions. Strains and plasmids used in this study are listed in **Supplementary Data 1**, and oligonucleotides are listed in **Supplementary Data 2**. Detailed information on part design, plasmid maps and corresponding experimental data for each construct are available via http://biofab.org/data/.

All plasmid manipulations were performed using standard molecular biology techniques[51]. All enzymes used for plasmid manipulations were obtained from New England Biolabs (NEB), and oligonucleotides were received from Integrated DNA Technologies (IDT). *E. coli* strain BW25113 was used for plasmid construction purposes and for fluorescence measurements (unless specified). All strains were grown in MOPS EZ Rich Medium (Teknova) supplemented with 50 μg/ml kanamycin (kan) at 37 °C, shaken at 900 r.p.m. All of the experiments were conducted in triplicate (biological replicates).

Plasmid library construction. The randomized bicistronic design (BCD) library, randomized promoter library (RPL), modular promoter library (MPL), combinatorial monocistronic design (MCD)–gene of interest (GOI) library, BCD-GOI library and promoter-BCD library were assembled on medium-copy vectors derived from pFAB217 (with the reporter *sfgfp*[52], termed *gfp* hereafter) and pFAB216 (with the reporter *mrfp1* (ref. 53), termed *rfp* hereafter). Both pFAB217 and pFAB216 were derived from the same backbone vector pBbA2k-RFP[54] (p15A replication origin, kan resistance) by replacing the Ptet promoter and *tetR* gene with a defined sequence context including the Ptrc* promoter and Bujard RBS region[55] (for further details on the neighboring sequence context, see "Design of an expression operating unit" below and the plasmid maps at http://biofab.org/data/) preceding either the reporter gene *gfp* (in pFAB217) or *rfp* (in pFAB216) (**Supplementary Figs. 28–31**).

All PCR amplifications were carried out with high-fidelity Phusion DNA polymerase (NEB, manufacturer's instructions). The primers used for vector amplification or for preparing an annealed product were phosphorylated using polynucleotide kinase in T4 DNA ligase buffer at 37 °C for 1 h and heat inactivated at 65 °C for 30 min.

Design of an expression operating unit. Both vectors pFAB217 and pFAB216 used to construct various backbone vectors (**Supplementary Figs. 28–31**) for the combinatorial libraries presented in this work have a defined microcontext, which we term as an 'expression operating unit' (EOU). The EOU comprises a minimal unit of genetic expression (expression cassette) and an additional flanking region that may play a role of insulation to EOU parts (**Supplementary Table 1**). The minimal unit of genetic expression is made up of a promoter with a defined transcription start site (Ptrc*, a constitutive promoter, −35 to +1), 5′ UTR[55], translation initiation element (BCD context, this work), a protein-coding region (for example, a reporter such as GFP or RFP) and a terminator (3′ UTR, dbi terminator[54]).

To provide functional insulation to the EOU from cryptic promoters, RBS-like regions, intrinsic terminators and AT-rich UP element–like features, we have introduced an additional upstream region composed of three-frame stop codons, an intrinsic terminator[56], a transcriptional pause site[57] and an insulator region[58] (see **Supplementary Table 1** for the entire EOU sequence). Here, the upstream and downstream terminators are designed and positioned to reduce the interactions between the EOU and the immediate genetic context. The EOU thus provides a standardized and well-defined context that insulates functional parts within the EOU from neighboring genetic contexts and provides a more reliable platform for characterization of parts. The use of standardized context thus helps in understanding and describing part performance relative to that of other parts. To facilitate joining of multiple EOUs (for example, to yield an expression operating system), these vectors have EcoRI-BglII sites upstream of the EOU and a XhoI-BamHI site downstream of the stop codon of a reporter, a configuration based on Bgl-Brick design[54]. The contribution and significance of the EOU design in insulating the functionality and functional composition of parts needs systematic characterization studies and has not been explored further.

Design and construction of the randomized BCD library. To generate the randomized BCD library, we first made the plasmid pFAB866 by amplifying the backbone vector pFAB217 (encoding the reporter GFP) using phosphorylated primers oFAB470 and oFAB472. These primers replace the 5′ UTR of pFAB217 with a bicistronic design with a translationally coupled second cistron encoding reporter GFP (**Supplementary Fig. 2**). The PCR-amplified vector backbone products were purified using Qiagen PCR purification kits, digested with DpnI (to remove the intact backbone vector), self-ligated using T4 DNA ligase enzyme and transformed into chemically competent BW25113 *E. coli* cells. Positive clones were then confirmed by sequencing, stored as glycerol stocks and used for preparing plasmid minipreps for further BCD library construction purposes.

For generating the randomized BCD library, pFAB866 was amplified using phosphorylated primers oFAB785 and oFAB786. The forward primer oFAB785 creates variants of the second SD of bicistronic design such that 3 nt upstream and downstream of the GGA motif of SD2 are randomized (NNNGGANNN; **Supplementary Fig. 2**). The PCR products were purified using Qiagen PCR purification kit, digested with DpnI, ligated using T4 DNA ligase, transformed into chemically competent BW25113 *E. coli* cells and grown overnight in selective LB agar medium (with kan). The next day, about 200 colonies were picked, and positive clones were confirmed by sequencing. We discarded mutants with STOP codons within cistron 1 in addition to deletion and insertion mutants within the leader peptide library to keep intact the −1 frame shift comprising the coupled BCD core. Positive clones were stored as glycerol stocks and assayed for bulk fluorescence on the plate reader (below).

Design and construction of the synthetic constitutive promoter library. We used two distinct approaches to engineer a diverse library of constitutive promoters for engineering gene expression in *E. coli*. In the first approach, we randomized the −10 and −35 motifs of a strong Ptrc* promoter (the asterisk indicates a promoter with no operator sites downstream of the transcription start site, −35 to +1 of the promoter) to generate an RPL, whereas in the second approach, an MPL was created by the combinatorial assembly of three modules of five well-characterized promoters of different strengths (**Fig. 1** and **Supplementary Figs. 20–22**). The sequences and plasmid maps for RPL and MPL members are listed with their corresponding promoter strengths at http://biofab.org/data/.

© 2013 Nature America, Inc. All rights reserved.

RPL. The RPL was created by randomizing the −10 motif (N*TA*NNN*T*N) or the −35 motif (N*TT*NNNN) or both the −10 and −35 motifs of a strong Ptrc* constitutive promoter (*TTGACA*ATTAATCATCCGGCTCG*TATAAT*GTGTGG**A**; consensus motifs are italicized, and bold 'A' is the transcription start site[54,59]) (**Supplementary Fig. 20**). This randomization strategy retains the most conserved and functionally important bases[37,60–62], with the expectation that it may alleviate the bias toward generating too many weak promoters.

To generate the RPL, we used plasmid pFAB217, which comprises the Ptrc* promoter and the Bujard RBS (ACAATTCATTAAA*GAGGAGA*AAGGTACC)[55] to drive the expression of the GFP reporter within EOU architecture. To randomize the −10 motif, we amplified pFAB217 using phosphorylated primers oFAB178 and oFAB177. The forward primer oFAB178 creates variants of the −10 motif (N*TA*NNN*T*N), and the reverse primer oFAB177 retains the consensus −35 motif of the Ptrc* promoter. To randomize the −35, the plasmid pFAB217 was amplified using phosphorylated primers oFAB176 and oFAB179. The forward primer oFAB176 retains the consensus −10 motif of the Ptrc* promoter, and the reverse primer oFAB179 creates variants of the −35 motif (N*TT*NNNN). The phosphorylated primers oFAB178 and oFAB179 were used to randomize both the −10 motif and the −35 motif.

The PCR products were purified using Qiagen PCR purification kit and digested with DpnI to remove the intact backbone vector. PCR products were then ligated using T4 DNA ligase, transformed into chemically competent BW25113 *E. coli* cells and grown overnight in selective LB agar medium (with kan) on three large QTray plates. The next day, about 2,000 colonies were picked from all three transformation plates (in total) and grown overnight (~16 h, at 37 °C, 900 r.p.m. on an Inforys shaker) in 500 μl MOPS EZ Rich +kan medium in 96–deep-well plates sealed with a breathable membrane. The following day, 250 μl of overnight culture was stored as a presequencing glycerol stock (250 μl overnight culture + 250 μl of 30% sterile glycerol), and the remaining 150 μl of the overnight culture was subjected to the microplate end-point assay to measure growth (optical density, $OD_{600\ nm}$) and fluorescence (relative fluorescence units or RFU) at an excitation of 481 nm and emission of 507 nm for GFP in a multimode microplate reader-incubator-shaker Synergy-2 (BioTek Instruments). With these preliminary promoter activity results, all promoters were grouped into ten bins of different strengths, and about 300 overnight cultures (showing a wide range of activity) were sent for sequencing.

The sequencing was performed on PCR product (using primers soFAB1 and soFAB8) comprising the cloning region using primers soFAB36 and soFAB37. Constructs with mutations in −10 and/or −35 motifs and single or double base deletion or addition in the spacer region were considered as positive clones, and constructs with mutations elsewhere on the plasmid were discarded from the library. The positive clones were stored as glycerol stocks and assayed for growth, bulk and single-cell fluorescence (see below).

MPL. The MPL was engineered by combinatorial assembly of three modules originating from five promoters of various strengths (with a known +1 transcription start site) to yield a total of 125 modular promoters. Here, one of the main objectives was to construct a synthetic promoter library made up of modules and key elements from different-strength promoters such that we obtain insight on how variation of different promoter elements (UP elements, −35 motif, spacer, −10 motif, discriminator region downstream of −10 motif to +1) affects promoter strength[37,63–65].

We used the strong T7A1 promoter[37], Ptrc promoter[54,59] and T5N25 promoter[66] and the weaker NM535 series[67] and U56D46 version of the pRM promoter series[68] as parental sequences for the MPL (**Supplementary Fig. 21**). The sequence of these five promoters was divided into three modules comprising (i) UP element and −35 motif, (ii) spacer region and (iii) −10 and spacer of −10 to +1. We then determined the promoter sequence of all 125 sequence combinations using an in-house–written Python script. We used a modified Golden Gate method[15,69] to assemble the promoters, using annealed oligonucleotides, into a restriction-digested plasmid.

To build the MPL, we first made plasmid pFAB517 by amplifying the backbone vector pFAB217 (encoding the reporter GFP) using phosphorylated primers oFAB124 and oFAB125. These primers replace the promoter Ptrc* in pFAB217 with type II restriction enzyme BsaI recognition sites on either strand of the vector such that after a post-restriction digestion of the ligated PCR products, we obtain appropriately compatible overhangs to clone promoter inserts. The PCR-amplified vector backbone products were purified using Qiagen PCR purification kits, digested with DpnI, self-ligated using T4 DNA ligase enzyme and transformed into chemically competent BW25113 *E. coli* cells. Positive clones were then confirmed by sequencing (using primers soFAB1 and soFAB8) and stored as glycerol stocks. Plasmid minipreps were prepared and used for further MPL construction purposes. Minipreps of these backbone vectors were then digested with BsaI enzyme (37 °C, overnight (> 16 h)), dephosphorylated, gel-purified (Qiagen) and used for assembling the promoter library.

To prepare the promoter elements as inserts for building the MPL, we designed 125 forward and 125 reverse oligonucleotides such that they can be annealed together and their overhangs are compatible with the restriction-digested backbone vector pFAB517. The forward and reverse oligonucleotides used for annealing the promoter parts are listed in **Supplementary Data 2**. For further details on assembling annealed parts in restriction digested vector see "Assembling the combinatorial libraries" (below). The positive clones were stored as glycerol stocks and assayed for growth, bulk and single-cell fluorescence (**Supplementary Fig. 21**; below).

Combinatorial libraries. For constructing the BCD:GOI, MCD:GOI and promoter:BCD combinatorial libraries, a modified Golden Gate method[15,69] was used to comply with the assembly of smaller parts or inserts (promoter, MCD and BCD). This type II endonuclease–mediated assembly method allows a scare-less and multipart assembly.

Construction of backbone vectors. To prepare the backbone vector for cloning of combinatorial libraries, phosphorylated forward and reverse oligonucleotides were used to PCR-amplify vectors pFAB217 and pFAB216. The forward and reverse primers introduce type II restriction enzyme BsaI recognition sites on either strand of the vector such that after a post-restriction digestion of the ligated PCR products, we obtain appropriately compatible overhangs to clone inserts (promoter, BCD, MCD, GOI or linkers). The PCR-amplified and purified vector products were then ligated,

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.2404

transformed into chemically competent *E. coli* DH10B cells and grown overnight on selective medium. Positive clones were then confirmed by PCR-amplifying and sequencing of the ligated region using sequencing primers soFAB1 and soFAB8. The overnight cultures of positive clones were stored in glycerol stocks as explained in the above section. The minipreps of these backbone vectors were then digested with BsaI enzyme (37 °C, overnight (>16 h)), dephosphorylated, gel-purified (Qiagen) and used for assembling combinatorial libraries. The six main backbone vectors pFAB870, pFAB871, pFAB1177, pFAB1178, pFAB1781 and pFAB1782 were constructed as described below for building combinatorial libraries reported in this work (**Supplementary Fig. 28**).

The backbone vector pFAB870 was constructed by amplifying pFAB217 using phosphorylated oFAB625 and oFAB584 primers and ligating the PCR product. The vector pFAB870 was used for cloning the combinatorial library of BCD and various GOI (either 36-nt or full-length) contexts fused to GFP.

The backbone vector pFAB871 was constructed by amplifying pFAB216 using phosphorylated oFAB626 and oFAB584 primers and ligating the PCR product. The vector pFAB871 was used for cloning the combinatorial library of BCD and various GOI (either 36-nt or full-length) contexts fused to RFP.

The backbone vector pFAB1177 was constructed by amplifying pFAB217 using phosphorylated oFAB950 and oFAB584 primers and ligating the PCR product. The vector pFAB1177 was used for cloning the combinatorial library of BCD fused to GFP.

The backbone vector pFAB1178 was constructed by amplifying pFAB216 using phosphorylated oFAB951 and oFAB584 primers and ligating the PCR product. The vector pFAB1178 was used for cloning the combinatorial library of BCD fused to RFP.

The backbone vector pFAB1782 was constructed by amplifying pFAB217 using phosphorylated oFAB950 and oFAB125 primers and ligating the PCR product. The vector pFAB1782 was used for cloning the combinatorial library of promoters and BCDs translationally fused to GFP.

The backbone vector pFAB1781 was constructed by amplifying pFAB216 using phosphorylated oFAB951 and oFAB125 primers and ligating the PCR product. The vector pFAB1781 was used for cloning the combinatorial library of promoters and BCDs translationally fused to RFP.

Preparation of inserts for constructing combinatorial libraries. To prepare the basic transcription and translation elements as inserts for building combinatorial libraries, we first phosphorylated and then annealed the forward and reverse oligonucleotides (by mixing 5 µl of 100 µM of forward and reverse primers with 90 µl of sterile water, incubating at 95 °C for 3 min and cooling at room temperature for 30 min). These annealed inserts were then diluted with sterile water such that the concentration was equivalent to that of the digested and purified vector. The sequences for a subset of all BCD variants, MCD variants, constitutive and inducible promoters and GOI regions and for a linker region and an EOU sequence are listed in **Supplementary Table 1**. Additional plasmid sequence and activity details are presented at http://biofab.org/data/. The forward and reverse oligonucleotides used for annealing the parts are listed in **Supplementary Data 2**.

Constitutive and inducible promoters. The 14 constitutive promoters used in promoter:BCD combinatorial library were chosen from a collection of synthetic constitutive promoters (see above: "Design and construction of the synthetic constitutive promoter library"). These promoters are variable in length (though they maintain a defined putative +1 mRNA start site) and have a wide range of promoter activities. To make the combinatorial assembly of these promoter parts with BCD parts easy to scale, we chose the promoters with the same spacer region between the −10 motif and the putative transcription start site (all promoter sequences used here are given in **Supplementary Table 1**). In addition to 14 constitutive promoters, a consensus 23-base-pair phage T7 promoter (TAATACGACTCACTATAGGGAGA) was chosen to test whether BCDs retain their functional reliability with T7 RNA Polymerase.

To test the functional reliability of BCDs with regulated promoters, we chose constitutive promoters of different strengths from the promoter libraries and replaced the promoter spacer region between the −35 and −10 motifs with LacI or TetR operator sequences[55] (**Supplementary Table 1**). Performance of one BCD (BCD2, apFAB682) with ten LacI- and nine TetR-regulated different-strength promoters (**Supplementary Table 1**) is shown in **Supplementary Figure 27**. These results demonstrate that inducible promoters retain their function when used in combination with BCD elements.

Bicistronic designs. Twenty-two BCDs having a wide dynamic range of translation initiation activity (used in the BCD:GOI combinatorial library and promoter:BCD combinatorial library) were all derived from the randomized BCD library presented in **Supplementary Figure 2** and are given in detail via **Supplementary Table 1**. Because all BCDs used in this work are ~80 nt in length, for easy handling, lower cost and improved quality of oligo synthesis, we decided to separate the BCDs into two parts. Part 1 is the invariable region of the BCD (surrounding RBS1), and part 2 is the variable region of BCD (surrounding RBS2). This design permits the use of the same part 1 for all assemblies in BCD:GOI combinations and P:BCD combinations except for a few special control cases, which use a different part 1 for assembling combinatorial libraries with the 22 sequence- and activity-distinct part 2 components (**Supplementary Data 2**). These are for (i) BCD:GOI and P:BCD combinatorial libraries (part 1 oligos oFAB979 and oFAB980); (ii) BCDs with early stop codons in the first cistron—in this case, we replaced the GUA_6 (valine) codon, a sixth codon of the first cistron with a UAA stop codon (part 1 oligos oFAB1638 and oFAB1639); (iii) BCDs with rare codons in the first cistron—three part 1 variants were designed by inserting different rare codons in the first cistron: (a) AGG_6 (arginine codon) replacing GUA_6 (valine) codon, a sixth codon of the first cistron (part 1 oligos oFAB1632 and oFAB1635); (b) AGG_4 (arginine codon) replacing ATT_4 (isoleucine) codon, a fourth codon of the first cistron and GGA_6 (glycine codon) replacing GUA_6 (valine) codon, a sixth codon of the first cistron (part 1 oligos oFAB1633 and oFAB1636); and (c) CGG_6 (arginine codon) replacing GUA_6 (valine) codon, a sixth codon of first cistron, and CTA_4 (leucine codon) replacing ATT_4 (isoleucine) codon, a fourth codon of the first cistron (part 1 oligos oFAB1634 and oFAB1637); (iv) BCD backbones with an inactive first SD (Null-SD1) motif (part 1 oligos oFAB981 and oFAB982)—to inactivate the SD site upstream of the first cistron, we replaced the native AAAGGAGAU motif with AACCUCCAU;

© 2013 Nature America, Inc. All rights reserved.

and (v) promoter T7:BCD combinations (part 1 oligos oFAB1361 and oFAB980)—as the sequence around the T7 transcription start site is different from the synthetic promoters used in this work, we designed compatible part1 for cloning part 2 BCDs.

Monocistronic designs. Twenty-two MCDs were assembled by annealing phosphorylated forward and reverse oligonucleotides (**Supplementary Data 2**). These MCDs have the same context around the RBS2 (that is, SD2) region as BCDs and yield a direct comparison of translation initiation around this RBS in the absence of translation from upstream RBS1 (that is, SD1).

Genes of interest. To test the reliability of BCDs, as compared to the standard 5′ UTRs (that is, MCDs), in initiating the translation of a sequence-independent coding region, we chose eight sequence-independent GOIs (**Supplementary Fig. 3**). These include *lacI* (EG10525, *E. coli* K-12), *araC* (EG10054, *E. coli* K-12), *gfp*[52], *rfp*[53], *tetR*[70] and a penicillin (cephalosporin) acylase gene (M18278)[71] from *Pseudomonas* sp. strain SE83, a codon-optimized putative cellulase (AAY81158)[72] gene from *Sulfolobus acidocaldarius* DSM 639 and a codon-optimized phosphomevalonate kinase gene from *Saccharomyces cerevisiae*[73,74]. The choice of these candidate genes was based on their sequence independence with each other, utility and importance in the ongoing in-house projects.

The sequence context of 36 nt from each N terminus of the various GOIs fused to the second codon of either *gfp* or *rfp* reporter gene (yielding total 14 chimeric reporter GOIs) is listed in **Supplementary Table 1** and was assembled by annealing phosphorylated forward and reverse oligonucleotides (**Supplementary Data 2**). For preparation of the full-length *tetR* gene as a GOI insert, we used primers oFAB1347 and oFAB1239 to PCR-amplify the *tetR* gene from the vector VKM81. The forward and reverse primers introduce BsaI recognition sites onto the N- and C-terminal ends of the PCR product such that a post-restriction digestion of the PCR product gives appropriate overhangs to clone into the digested backbone vector along with additional inserts, such as a linker region (see **Supplementary Figs. 29** and **30**).

To examine a BCD's capacity to overcome the impact of hairpin formation spanning the junction of SD2 and the GOI initiation codon on translation initiation, we designed two special GOIs. These GOIs have sequence complementarity to a strong SD2 motif of BCDs (UAAGGAGGU) such that the mRNA structure predictions indicated a stronger hairpin formation around the translation initiation region (**Supplementary Fig. 1**). These two GOIs have the same 36-nt *tetR* gene as the backbone with 9 nt downstream of the start codons mutated such that there is potential (variable-strength) hairpin formation between SD2 and the GOI start codon region (**Supplementary Table 1** and **Supplementary Fig. 1**).

Linker region. The full-length *tetR-gfp* fusion includes a linker region between the TetR and GFP coding regions. This glycine-rich linker region also includes a Tev protease site, which can be cleaved if needed (**Supplementary Table 1**).

Sequence-independent BCDs. To test the generality of the BCD across different GOIs, we assembled sequence-independent BCDs as listed in **Supplementary Table 1** (**Supplementary Figs. 18** and **19**). In these constructs, we used seven RBS2 regions from intercistronic regions of operons (BCD2 (this work), LeuL, HisB-H junction, TrpB-A junction, LeuA-B junction, HisH-A junction and HisC-B junction)[21–26,75,76], whose junctions have overlapping stop-start codon motifs (TAATG) and have SD2 motifs upstream of the stop-start junction (**Supplementary Fig. 18**). The six RBS1 regions were chosen from different 5′ UTRs[45,75,77,78], and the SD1 motif of the RBS1 region was mutated to a consensus SD motif so that translation initiation of the first cistron was not limiting. These sequence-independent BCDs were cloned upstream of the *gfp* reporter as explained earlier (on restriction-digested vector pFAB1177) and characterized by measuring fluorescence. Several representative BCD candidates were chosen for further characterization by replacing the SD2 motif with different strength variants presented in **Supplementary Table 1** and cloned upstream of *gfp* and *rfp* reporter genes. The data shown in **Supplementary Figures 18** and **19** demonstrated that BCD variants retain their functionality across different GOIs and are generalizable. These sequence-independent BCDs are useful in constructing heterologous pathways or genome-scale engineering efforts. Further studies on sequence-independent BCDs are essential to understand any impacts of overproduction of different peptides on cellular factors or growth.

Assembling the combinatorial libraries. The general process for assembling the combinatorial libraries is shown as a schematic in **Supplementary Figure 32**. All of the cloning steps including phosphorylation of oligonucleotides, annealing of phosphorylated oligonucleotides, dilution of annealed products, ligation of annealed products (inserts) with cut vector backbone, incubation of ligation reactions and transformation were carried out in 96-well PCR plates.

All ligation reactions were 10 μl in total volume and made up of 1 μl of each of the annealed parts (~10 ng/μl), 1 μl of digested and pure vector backbone (~10 ng/μl), 1 μl of ligase enzyme, and appropriate volumes of ligase buffer and sterile water to make up the total volume. The ligation reaction was run for 30 min at room temperature (20–22 °C) using concentrated T4 ligase enzyme and then moved onto ice. The ligation reaction was then incubated with 50 μl of chemically competent *E. coli* cells (BW25113 (ref. 79) and DH5αZ1 (ref. 55) for *E. coli* RNAP or BL21(DE3) for T7 RNAP) in 96-well plates (in-house–prepared BW25113; DH5αZ1cells and BL21 from NEB) for 30 min on ice. The transformation step was performed with heat shock at 42 °C for 90 s in a PCR machine, and then the plates were moved onto ice for 2 min before 100 μl of sterile SOC medium was added. The transformation reaction in 96-well plates was then incubated at 37 °C for 1 h with 900-r.p.m. shaking.

We used the vented QTray with 48-well dividers (Genetix, cat. no. X6029) for plating 35 μl of transformation reaction (leftover reaction mix was stored at 4 °C overnight and discarded the next day for transformations that worked) on solid LB agar plates with kan. Contents of each 96-well plate transformation reaction were plated out on two 48-well QTrays with LB agar +kan. To spread 35 μl of transformant reaction evenly across each of 48 wells, we used 10–15 sterile glass beads per well and stirred gently (with the lid on) to avoid the mix-up of beads between wells. After we removed the beads (by quickly turning the plates upside down and collecting beads on plate lids), the plates were allowed to dry

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.2404

and were incubated overnight at 37 °C. The next day, individual transformant colonies were picked for sequence confirmation and for preparing glycerol stocks. Two colonies per transformant were suspended in 50 µl EB buffer (pH = 8) in a 96-well plate. From this colony suspension, 25 µl was sent to a sequencing service, and the leftover suspension was used to inoculate 250 µl of LB +kan medium in a 96-well plate. The next day, the overnight culture plate was stored as a presequencing glycerol stock (250 µl overnight culture + 250 µl of 30% sterile glycerol) until the sequencing results were obtained and analyzed. The sequencing was performed by PCR-amplifying the cloning region using primers (soFAB1 and soFAB8), and sequencing was done using primers soFAB36 and soFAB37.

Once the sequencing results were obtained, the correct clones (from the presequencing glycerol stocks) were used to inoculate fresh LB +kan medium in 96–deep-well plates and were grown overnight and stored as main glycerol stocks.

Construction of tRNA complementation plasmid. To study the impact of rare codons in the leader cistron of the BCD on downstream gene expression, we constructed 12 plasmids with three different rare codons (at the fourth and sixth codons within the leader cistron) in the context of four different-strength BCDs (**Fig. 3d**). As a control, we also inserted early stop codons (at the sixth codon of the leader cistron) in the context of four different-strength BCDs. To study the impact of complementing the tRNA for rare codons on the gene expression and rank order of BCDs, we chose the plasmid pRARE2 (Novagen), which contains various tRNA genes for the following rare codons in *E. coli*: AGA, AGG (Arg), GGA (Gly), AUA (Ile), CUA (Leu), CCC (Pro). This plasmid has a chloramphenicol (Cam) resistance cassette and P15A replication origin, and all of the tRNA genes have their endogenous promoters. Because of the incompatibility between the plasmid pRARE2 and all the constructs reported in the present work (both have P15A replication origins), we decided to replace the replication origin of pRARE2 with a ColE1 origin.

As the plasmid sequence of pRARE2 is proprietary and unavailable to users, we designed various primers to sequence the region around the P15A replication origin (oFAB1611, oFAB1612 and oFAB1613) and found specific restriction digestion sites for NheI and XbaI enzymes around the replication origin. We prepared plasmid DNA for pRARE2 from *E. coli* Rosetta 2 (DE3), digested it with NheI and XbaI (NEB) enzymes and gel-purified the digested plasmid. The replication origin ColE1 was PCR-amplified from the plasmid VKM74 using primers oFAB1624 and oFAB1625. The forward and reverse primers introduce NheI and XbaI digestion sites such that after digestion with both enzymes, the PCR product is compatible for ligation with the NheI-XbaI–digested pRARE2. The ligation of the NheI-XbaI–digested PCR product and pRARE2 vector and the transformation of ligation reaction were done according to the standard procedure. The positive clones were confirmed by sequencing with primers oFAB1611 and oFAB1612. Once the sequence was confirmed, we performed miniprep on the pRARE2-ColE1 plasmid (pFAB4526), transformed it into assay strain BW25113 and subsequently stored it as a glycerol stock.

To study the impact of overexpression of tRNA genes for rare codons and its effect on the rank order of BCDs (with rare codons and an early stop codon in the leader cistron), we cotransformed pFAB4526 with BCD constructs having rare codons in the leader cistron (**Supplementary Data 1**) driving the expression of GFP and RFP. Transformants were then selected and grown on kan and cam selection medium for assay purposes and for storing the glycerol stocks.

***In vivo* assays using the flow cytometer.** Assay strains were stored as main glycerol stocks in 96–deep-well plates (2 ml) and in smaller aliquots of 50 µl in 96-well sterile PCR plates as working stocks. Cultures were grown in 2 ml 96–deep-well plates containing 400 µl of MOPS EZ Rich Medium (Teknova, cat. no. M2105) with appropriate antibiotics and inoculated with 3 µl from thawed glycerol stocks. Cultures were grown overnight (~16 h) in 96-, U-shaped-, 2-ml-well plates covered with sterile breathable sealing film at 37 °C with shaking at 900 r.p.m. on a Multitron shaker (Inforys-HT).

For microplate end-point assays (to measure the optical density and fluorescence), the overnight cultures were diluted 1:50 into a final volume of 400 µl fresh MOPS EZ Rich Medium with appropriate antibiotics in 1-ml-deep–well plates and grown for 2 h at 37 °C with shaking at 900 r.p.m. on a Multitron shaker. Samples were collected (150 µl in clear-bottom black plates) to measure growth (optical density, $OD_{600\ nm}$) and fluorescence (RFU; excitation at 481 nm and emission at 507 nm for GFP; excitation at 560 nm and emission at 650 nm for RFP) in a multimode microplate reader-incubator-shaker Synergy-2 (BioTek Instruments). Repeated assays showed that we were sampling the cultures at OD_{600} of 0.3–0.5 and that these cultures were in the exponential growth phase. All experiments were repeated at least three times. Gen5 software for the BioTek plate reader was used for data acquisition, and further data analysis was performed using MATLAB software (MathWorks) with in-house–developed scripts.

For the flow cytometer assays, the overnight cultures of BW25113 cells with plasmid libraries were diluted 1:50 into a final volume of 200 µl fresh MOPS EZ Rich Medium with appropriate antibiotics in 1-ml-deep–well plates and grown for 2 h (to exponential phase with OD_{600} in the range of 0.3–0.5 in the microplate reader) at 37 °C with shaking at 900 r.p.m. on a Multitron shaker.

For constructs encoding a T7 promoter, the overnight cultures of BL21 (DE3) with plasmid libraries were diluted 1:50 in to a final volume of 200 µl fresh MOPS EZ Rich Medium with appropriate antibiotics and 0.4 mM IPTG (to induce T7 RNAP expression from wild-type *lac* promoter on chromosome) in 1-ml-deep–well plates and grown for 2 h (to exponential phase with OD_{600} in the range of 0.3–0.5 in the microplate reader) at 37 °C with shaking at 900 r.p.m. on a Multitron shaker.

For inducible promoter:BCD combinations, the overnight cultures of DH5αZ1 (wherein LacI and TetR were constitutively expressed from the bacterial chromosome) with plasmid libraries were diluted 1:50 in to a final volume of 200 µl fresh MOPS EZ Rich Medium with appropriate antibiotics and 1 mM IPTG or 100 ng/ml aTC in 1-ml-deep–well plates and grown for 2 h (to exponential phase with OD_{600} in the range of 0.3–0.5 in the microplate reader) at 37 °C with shaking at 900 r.p.m. on a Multitron shaker.

Cultures at exponential phase were diluted 1:2,000 in chilled and filtered PBS (Gibco, pH 7.4) containing 500 µg/ml streptomycin in chilled 96-well clear plates (Costar) and immediately subjected to flow cytometer analysis. We used a Guava EasyCyte flow cytometer (EMD Millipore) equipped with autosampling capabilities

© 2013 Nature America, Inc. All rights reserved.

and paired dual blue (488-nm, 75-mW) and green (532-nm, 40-mW) laser excitation with two customized filter options for emission detection of 510/20 for GFP and 610/20 for RFP, respectively. During the assay, the sample concentration was kept below 500 cells per µl, and samples were run on a high flow rate (1.18 µl/s) until 2,000 cells (with a range of 60–300 events per µl) had been collected within small forward- and side-scatter gates. Guavasoft software was used for data acquisition, and the resulting FCS files were further analyzed using in-house–developed R scripts[15]. The fluorescence-per-cell values for each GOI construct were $\log_2$-transformed and then mean-normalized for comparative analysis of fluorescence from sequence-distinct GOI fusions.

Absolute and mean-normalized expression. Absolute observed fluorescence values for all genes tested depended on the selected fluorophore and the specific 36-nt coding-sequence leader. To visually compare the rank-ordered activities of 5′ UTRs encoding MCDs and BCDs across various GOIs, we estimated mean-normalized expression levels from absolute expression data, wherein we divided absolute expression values for any given 5′ UTR:GOI combination by the average of all absolute expression values for a given GOI and 5′ UTR design (for example, the average for a given gene across all MCD:GOI absolute expression levels) (**Fig. 2a**,**b**,**e** and **Supplementary Figs. 5** and **7**).

Sequence-identity calculations. Global pairwise alignment of the 36-nt sequences were computed using the emboss implementation of the Needleman and Wunch algorithm[80,81] with default parameters. Percentage identities were calculated from these alignments as the number of matching nucleotides divided by 36. The average identity between GOIs was 27% with an s.d. of 23%. These values are comparable to what one would expect by chance. **Supplementary Figure 3** shows the percentage identities for different GOIs used in this work.

Free-energy calculations of mRNA folding at the MCD:GOI or BCD:GOI junction. To understand the potential for forming stable inhibitory structures between different chosen GOIs with 5′ UTRs, we used UNAfold software[82] to predict the minimum-folding-energy structure conformation. We considered the junction region to comprise between the positions −26 and +37 with respect to the translation start site. These boundaries were selected on the basis of the size of the monocistronic 5′ UTR (MCD) and 36-nt region of the GOI, respectively. The predicted minimum free-energy calculations depicted a wide diversity in the stability of mRNA structures formed at the translation initiation region of GOIs (**Supplementary Fig. 3**).

Hybridization-energy calculations for the SD2 variant–16S RNA duplexes. To evaluate the affinity between SD2 sequences (in BCD and in MCD) and the SD-complementary region from the 16S rRNA (ACCTCCTTA), we used UNAfold software[82] to calculate the hybridization energy for each resulting RNA duplex. We considered the region spanning from positions −26 to −1 with respect to the translation start site. As this region is the same for both BCD and MCD constructs, we can use these free-energy calculations and correlate with the fluorescence measurements from fusion reporters for both MCD and BCD constructs (**Fig. 2f** and **Supplementary Figs. 11**, **12** and **15**).

Use of RBS Calculator to predict the ΔG_{total}. The current version of the RBS Calculator software[10] was downloaded from https://github.com/hsalis/ribosome-binding-site-calculator/ (download date: 3 June 2012). We wrote a script in the Ruby programming language to automate the analysis and used the calculator to estimate the total ΔG as defined in ref. 10. We used 5′ UTR sequences spanning from 27 nt upstream to 33 nt downstream of the start codons of *gfp*, *rfp*, *lacI*, *tetR*, *araC* and PMK, PA or cellulase gene fusions (**Supplementary Figs. 13–15**).

ANOVA models for MCD:GOI and BCD:GOI combinatorial data sets. To understand the contribution and coupling between translation elements (i.e., MCD and BCD) and the GOI on the overall gene expression, we performed ANOVA as reported in ref. 15. Briefly, we performed ANOVA on the following linear model using fluorescence data from chimeric GFP fusions

$$\log(\text{Fluorescence}_{ij}) = \alpha + U_i + \text{GOI}_j + (U:\text{GOI})_{ij} + \varepsilon_{ijk} \quad \text{for } i = (1-22);\ j = (1-8) \qquad (1)$$

where Fluorescence_{ij} is the fluorescent output signal measured from a genetic construct comprising a translation element, U_i, and a gene of interest, GOI_j. U:GOI_{ij} represents any interaction between the ith translational element and jth gene of interest, α is the overall average signal, and the term ε_{ijk} represents the error term for each particular U:GOI combination. In this approach, we assume that log(gene expression) is a linear function of different factors and their interactions, whereas each factor is an abstraction of the complex biophysical functions encoded at the sequence level. For example, U_i captures contributions due to ribosome binding and mRNA stabilization (codon usage and translation elongation in the case of BCD), which results in differential rates of translation initiation and transcript degradation, whereas its interaction term with GOI, $(U\text{:GOI})_{ij}$ describes the impact of the GOI on each U's translation initiation rate (for example, due to inhibitory mRNA structures or modification of transcript stability). The factor GOI_j defines the intrinsic differences in translation elongation property of codons that are coding this region (translation pause, codon effects and folding of polypeptide), protein degradation and fluorescence intensity itself. The analysis outputs are presented in **Supplementary Figures 9** and **10**.

ANOVA models for promoter:BCD:GOI combinatorial data sets. To understand the contribution and coupling between a transcriptional (P) element, translation (U) element and the fluorescent reporter on overall gene expression, we performed ANOVA on the following linear model

$$\log(\text{Fluorescence}_{ijk}) = \alpha + P_i + U_j + \text{GOI}_k + (P:U)_{ij} + (U:\text{GOI})_{ik} + (P:\text{GOI})_{jk} + (P:U:\text{GOI})_{ijk} + \varepsilon_{ijk} \quad \text{for } i = (1-14);\ j = (1-22);\ k = (1,2) \qquad (2)$$

where $\text{Fluorescence}_{ijk}$ is the fluorescent output signal measured from a genetic construct comprising a transcriptional element i, a translation element j and a reporter k. $(P\text{:}U)_{ij}$ represents the effect of any interaction between the ith transcriptional element and jth translational element; $(P\text{:GOI})_{ik}$ represents the effect of interaction between the ith transcriptional element and kth

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.2404

reporter; $(U{:}\mathrm{GOI})_{jk}$ represents effects of any interactions between the jth translational element and kth reporter; $(P{:}U{:}\mathrm{GOI})_{ijk}$ represents the interaction between the ith transcriptional element, jth translational element and kth reporter; α is the overall average signal; and the term ε_{ijk} represents the error term for each particular combination. The analysis outputs are presented in **Supplementary Figure 26**.

ANOVA models: sum of squares and score calculations. The models described in equations (1) and (2) relate the Fluorescence (proxy for protein abundance) to the transcriptional and translational elements that comprise each genetic construct. Using three replicates of fluorescence, we performed ANOVA[83] on the linear models described above using the "anova" routine in R software (http://www.r-project.org/). ANOVA results are presented in **Figure 2c**,**d** and **Supplementary Figures 9**, **10** and **26**. To account for the differences in fluorescence intensities of reporter fusions, we normalized the data sets with their respective mean fluorescence for each GOI, thus disregarding the part of the variance arising from the GOI factor. The main effects (the primary scores for promoters, BCDs and GOI reporters) were directly retrieved from the ANOVA table of effects (accessed using the "model.tables" function in R) as explained elsewhere[15]. The integrated deviation of the main effect (secondary scores) for each element, resulting from its composition with different parts, was calculated as the s.e.m. of the appropriate interaction term effects as described in ref. 15 and is shown in **Supplementary Figures 9**, **10** and **26**.

Predictive regression model of promoter:BCD combinatorial library. A full-factorial ANOVA (linear model, equation (2)) modeling on the observed fluorescence from members of the promoter:BCD combinatorial library showed 98% of the total dynamic fluorescence range was due to encoded differences in the intrinsic activity of individual promoters and BCDs (**Fig. 4e**), and ~1% of the variance was explained by the element-element interaction (promoter:BCD and BCD:GOI). Given the high degree of explanatory power, and the independence of the elementary parts, we hypothesized that a regression model for predicting expression from the identity of a particular promoter and BCD trained on expression measurements of any given reporter could be used to predict the expression of another GOI using the same translation control elements. To do this we considered a simplified linear model with the GOI held constant (equation below).

$$\log_2(\mathrm{Expression})_{ij} = \beta_i(P_i) + \gamma_j(U_j) \qquad (3)$$

where β_i and γ_j are the strengths for the ith and jth promoter and translation element, respectively.

In this categorical regression model, each promoter and BCD is a separate object/variable that can be recoded within a matrix of 1s (for presence) and 0s (for absence) that serve as predictors, with log-transformed fluorescence values serving as the response variable and betas representing regression weights, where i varies from promoter 1 to 14 and j varies from BCD 1 to 22. To build a predictive regression model based on recoded predictors and experimentally observed GFP fluorescence values, we used the partial least-squares regression (PLSR) approach[84].

We used the Unscrambler X10 (CAMO software) for PLSR model (PLSR1) building and calculation of regression coefficients. All models were built by applying the standard data preprocessing procedures. To test whether the model was overfitting the data, tenfold cross-validation was performed. This cross-validated model explained ~96% of the variance in the fluorescence data (cross-validated R^2 of 0.96, r.m.s. error of 0.25 with two principal components). We used the cross-validated model trained on the experimental data set from the promoter:BCD:GFP combinatorial library (**Fig. 4b**) to predict the RFP fluorescence from the same combination of transcription and translation elements (**Fig. 4c**) as well as the expression of other GOI fusions from the BCD:GOI combinatorial library driven by promoter P14 and 22 BCDs (**Fig. 2b**). The model successfully predicted the RFP (with R^2 of 0.9) and other GOI expression data sets (with R^2 of 0.89; combined RFP and GOIs yielded an R^2 of 0.9 and r.m.s. error for prediction of 0.48). Note that the GOIs are expressed on a different vector series than the vector used for the promoter:BCD combinatorial library (**Supplementary Fig. 28**), which demonstrates the prediction reliability across DNA contexts. The predicted output results with deviations are shown in **Figure 4f**.

Expression probability calculations. The probability of observed expression falling within a factor of 2 of the predicted expression was determined using two separate methods. For each of the strains, the means of the mean-normalized $\log_2$ fluorescence values (observed) and the predicted values were calculated for a total of 440 pairs of observed and predicted values (**Fig. 4f**, see above). The absolute difference between the observed and predicted values was calculated.

For the first method, the percentage of absolute difference values <1 was empirically determined to be 93.86%.

For the second method, a histogram was generated using bin sizes (W) calculated according to Freedman and Diaconis[85] using the formula

$$W = 2\times(\mathrm{IQR})\times N^{-1/3} \qquad (4)$$

where N is the number of samples and IQR is the interquartile range, defined as the 75th percentile minus the 25th percentile. A Gaussian was fitted to the histogram, and the probability of the error between observed and predicted being less than or equal to a factor of 2 was determined using the formula

$$\mathrm{erf}(\log_2(2)/(\sigma\times\sqrt{2})) \qquad (5)$$

with $\sigma = 0.5437$ from the fitted Gaussian and 'erf' is shorthand for the error function. The result using this method is 93.41% of observations falling within a factor of 2 of the predicted values. The estimated ~87% error reduction reflects a decrease in reported expression level errors from 53% (ref. 10) to 7% (this work).

Data representation. The heat map representations and hierarchical clustering of combinatorial data sets were performed using Multiexperiment Viewer (MeV) software[86]. The sequence logos were generated using the WebLogo web-based application[87].

Selected statistics[83].

Coefficient of determination (R^2). We used R^2 to represent how well simple linear regression models fit various data sets and, thus, to what extent models can be used to predict future outcomes. The value of R^2 can range from 0 (poor fit) to 1 (perfect fit).

© 2013 Nature America, Inc. All rights reserved.

For example, we found that standardized promoter and BCD elements could be used to express GFP across a range of levels and developed a model predicting expression levels for other genes from the GFP data. We then found that the observed expression levels for other genes were well correlated to predictions made using only the GFP data ($R^2 = 0.9$, **Fig. 4f**).

Pearson's correlation coefficient (r). Also known as the 'sample correlation coefficient', we used *r* to represent the covariance of two variables divided by the product of each variable's s.d. The value of *r* can range from −1 to 1 and can thus be used to communicate the 'direction' of a correlation. For example, we observed a negative correlation between 16S rRNA + SD mRNA binding free energies and resulting protein expression levels (*r* = various values, **Fig. 2f**).

Spearman's rank correlation coefficient (rho). Also known as 'Spearman's rho', we used this nonparametric statistic measure to assess the extent to which the relationship between two variables can be represented via a monotonic function. The value of rho can range from −1 to 1 in representing inverse to direct correlation of rank orderings, respectively. For example, we found that the rank correlations for the activities of BCDs, when used across multiple GOIs, was much higher than when the same SD sequences were used within MCDs (**Fig. 2a**,**b** and **Supplementary Figs. 6** and **8**). Stated differently, we used rho to quantify nonparametrically to what extent BCDs improved preservation of rank ordering for translation initiation elements as compared to MCDs.

Variance. We used this statistic to quantify to what extent the intrinsic activities encoded by various genetic elements lead to unexpected differences in observed protein expression. For example, we found that MCDs led to much more widely varying expressed protein levels relative to the levels realized using BCDs (**Fig. 2a**,**b**).

© 2013 Nature America, Inc. All rights reserved.

51. Ausubel, F.M. *Short Protocols in Molecular Biology* 5th edn. (Wiley, New York, 2002).
52. 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).
53. Campbell, R.E. *et al.* A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. USA* **99**, 7877–7882 (2002).
54. Lee, T.S. *et al.* BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *J. Biol. Eng.* **5**, 12 (2011).
55. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I_1–I_2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
56. McDowell, J.C., Roberts, J.W., Jin, D.J. & Gross, C. Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. *Science* **266**, 822–825 (1994).
57. Kireeva, M.L. & Kashlev, M. Mechanism of sequence-specific pausing of bacterial RNA polymerase. *Proc. Natl. Acad. Sci. USA* **106**, 8900–8905 (2009).
58. Davis, J.H., Rubin, A.J. & Sauer, R.T. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res.* **39**, 1131–1141 (2011).
59. Brosius, J., Erfle, M. & Storella, J. Spacing of the −10 and −35 regions in the *tac* promoter. *J. Biol. Chem.* **260**, 3539–3541 (1985).
60. Saecker, R.M., Record, M.T. Jr. & Dehaseth, P.L. Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *J. Mol. Biol.* **412**, 754–771 (2011).
61. Gross, C.A. *et al.* The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 141–155 (1998).
62. Shultzaberger, R.K., Malashock, D.S., Kirsch, J.F. & Eisen, M.B. The fitness landscapes of *cis*-acting binding sites in different promoter and environmental contexts. *PLoS Genet.* **6**, e1001042 (2010).
63. Rhodius, V.A., Mutalik, V.K. & Gross, C.A. Predicting the strength of UP-elements and full-length *E. coli* σE promoters. *Nucleic Acids Res.* **40**, 2907–2924 (2012).
64. Rhodius, V.A. & Mutalik, V.K. Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, σ^E. *Proc. Natl. Acad. Sci. USA* **107**, 2854–2859 (2010).
65. Mutalik, V.K., Nonaka, G., Ades, S.E., Rhodius, V.A. & Gross, C.A. Promoter strength properties of the complete sigma E regulon of *Escherichia coli* and *Salmonella enterica*. *J. Bacteriol.* **191**, 7279–7287 (2009).
66. Bujard, H. *et al.* A T5 promoter-based transcription-translation system for the analysis of proteins *in vitro* and *in vivo*. *Methods Enzymol.* **155**, 416–433 (1987).
67. Miroslavova, N.S. & Busby, S.J. Investigations of the modular structure of bacterial promoters. *Biochem. Soc. Symp.* **73**, 1–10 (2006).
68. Szoke, P.A., Allen, T.L. & deHaseth, P.L. Promoter recognition by *Escherichia coli* RNA polymerase: effects of base substitutions in the -10 and -35 regions. *Biochemistry* **26**, 6188–6194 (1987).
69. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).
70. Postle, K., Nguyen, T.T. & Bertrand, K.P. Nucleotide sequence of the repressor gene of the TN10 tetracycline resistance determinant. *Nucleic Acids Res.* **12**, 4849–4863 (1984).
71. Matsuda, A., Toma, K. & Komatsu, K. Nucleotide sequences of the genes for two distinct cephalosporin acylases from a *Pseudomonas* strain. *J. Bacteriol.* **169**, 5821–5826 (1987).
72. Master, E.R., Zheng, Y., Storms, R., Tsang, A. & Powlowski, J. A xyloglucan-specific family 12 glycosyl hydrolase from *Aspergillus niger*: recombinant expression, purification and characterization. *Biochem. J.* **411**, 161–170 (2008).
73. Redding-Johanson, A.M. *et al.* Targeted proteomics for metabolic pathway optimization: application to terpene production. *Metab. Eng.* **13**, 194–203 (2011).
74. Martin, V.J., Pitera, D.J., Withers, S.T., Newman, J.D. & Keasling, J.D. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.* **21**, 796–802 (2003).
75. Blattner, F.R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
76. Gulevich, A.Y. *et al.* A new method for the construction of translationally coupled operons in a bacterial chromosome. *Mol. Biol.* **43**, 505–514 (2009).
77. Olins, P.O., Devine, C.S., Rangwala, S.H. & Kavka, K.S. The T7 phage gene 10 leader RNA, a ribosome-binding site that dramatically enhances the expression of foreign genes in *Escherichia coli*. *Gene* **73**, 227–235 (1988).
78. Olins, P.O. & Rangwala, S.H. A novel sequence element derived from bacteriophage T7 mRNA acts as an enhancer of translation of the *lacZ* gene in *Escherichia coli*. *J. Biol. Chem.* **264**, 16973–16976 (1989).
79. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
80. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
81. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
82. Markham, N.R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.* **453**, 3–31 (2008).
83. Wu, C.F.J. & Hamada, M.S. *Experiments: Planning, Analysis, and Optimization* 2nd edn. (Wiley, Hoboken, New Jersey, USA, 2009).
84. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
85. Freedman, D. & Diaconis, P. On the histogram as a density estimator: L2 theory. *Z Wahrscheinlichkeit* **57**, 453–476 (1981).
86. Saeed, A.I. *et al.* TM4 microarray software suite. *Methods Enzymol.* **411**, 134–193 (2006).
87. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

doi:10.1038/nmeth.2404