Predictable tuning of protein expression in bacteria

Mads T Bonde[1,2], Margit Pedersen[1,2], Michael S Klausen[1,2], Sheila I Jensen[1], Tune Wulff[1], Scott Harrison[1], Alex T Nielsen[1], Markus J Herrgård[1] & Morten O A Sommer[1]

We comprehensively assessed the contribution of the Shine-Dalgarno sequence to protein expression and used the data to develop EMOPEC (Empirical Model and Oligos for Protein Expression Changes; http://emopec.biosustain.dtu.dk). EMOPEC is a free tool that makes it possible to modulate the expression level of any *Escherichia coli* gene by changing only a few bases. Measured protein levels for 91% of our designed sequences were within twofold of the desired target level.

The ability to precisely modify gene expression levels is critical for constructing new functional structures such as cell factories and biological circuits, and it represents a key challenge for synthetic biology. Recent efforts have focused on measuring the activity of modular genetic parts in plasmid systems for better forward engineering[1–4]. One goal has been to establish a toolbox of standard components for transcription and translation initiation. For example, measuring the performance of approximately 500 standard genetic elements alone and in limited combinations was sufficient for predicting activity in larger circuits[1,2]. Alternatively, efficient screening of synthetic libraries, such as the expression of 12,563 combinations of common promoters and ribosome-binding sites, using fluorescence-activated cell sorting (FACS) and deep sequencing (Flow-seq) suggested that such approaches could be used in place of prediction or standardization[3].

Although plasmid-based strategies show great promise, several applications, including cell-factory engineering, rely on genomically encoded proteins for which the insertion of large elements and insulators may be problematic (for example, owing to unpredictable consequences of inserting insulators in operons, overlapping genes or the low efficiency of genomic replacement for larger DNA sequences[5,6]).

It is now possible to modulate the expression of multiple chromosomally encoded genes using multiplex automated genome engineering (MAGE)[7] and its derivative microarray MAGE[8]. Mining diverse cell libraries created by these methods enables the systematic optimization of metabolic pathways and strain selection[7,9]. However, these approaches are very inefficient at introducing larger modifications such as insulator elements[1,2,10].

In bacteria, initiation is the rate-limiting step of translation and a major determinant of overall protein expression[11]. In general, the process requires an initiation codon and an upstream Shine-Dalgarno (SD) sequence, which is often a variation of the AGGAGG consensus[12–14]. The complementarity of the SD sequence and the anti-SD sequence in 16S ribosomal RNA has a strong influence on translation initiation[12,14], making the SD sequence an ideal engineering target—changing even a few bases can lead to substantial changes in the translation level. However, knowledge of the relationship between SD sequence and protein expression has been limited to a subset of studied SD sequences and thermodynamic models; this has made forward engineering difficult because of confounding factors such as codon usage in the N-terminal protein sequence[15–17].

To map the relationship between SD sequence and protein level, we applied MAGE with randomized six-base oligomers (N6) to comprehensively mutate the SD sequence of a constitutive promoter driving the GFP gene integrated in the *E. coli* K12 MG1655 genome (Online Methods). Cells were segregated into 16 GFP-expression bins using FACS (**Fig. 1a**) followed by DNA extraction, SD-sequence amplification and sequencing from each bin (Flow-seq). From 12.5 million usable reads, we identified 4,066 (or 99.3%) of the 4,096 (4^6) possible SD sequences, 3,087 of which exceeded our cutoff of 50 reads for reliable quantification.

We observed low correlation between protein expression (Online Methods and **Supplementary Table 1**) and the calculated hybridization energy between 16S rRNA and SD sequences (**Fig. 1b**). Sequences with a medium to strong hybridization energy, in particular, spanned a wide range of expression levels. Sorting normalized SD-sequence read counts by mean protein expression showed that most sequences led to low expression (**Fig. 1c**; 65% of sequences produced 10% of maximal expression) and were distributed across multiple bins with a clear peak (**Fig. 1d**).

To validate expression estimates from Flow-seq, we isolated 106 clones across all 16 bins and measured GFP levels using a plate reader (**Supplementary Table 2**). Individually measured SD sequences were highly correlated with Flow-seq estimates ($R^2 = 0.95$, $P < 10^{-67}$, linear regression) across a 200-fold expression range (**Fig. 1e**).

We compared our experimental data to those from state-of-the-art computational models based on the binding energy between ribosome and mRNA, spacing between the start codon and the SD sequence, and other parameters[16]. We used RBS Calculator[15]

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark. [2]These authors contributed equally to this work. Correspondence should be addressed to M.O.A.S. (msom@bio.dtu.dk).

RECEIVED 13 MAY 2015; ACCEPTED 13 NOVEMBER 2015; PUBLISHED ONLINE 11 JANUARY 2016; DOI:10.1038/NMETH.3727

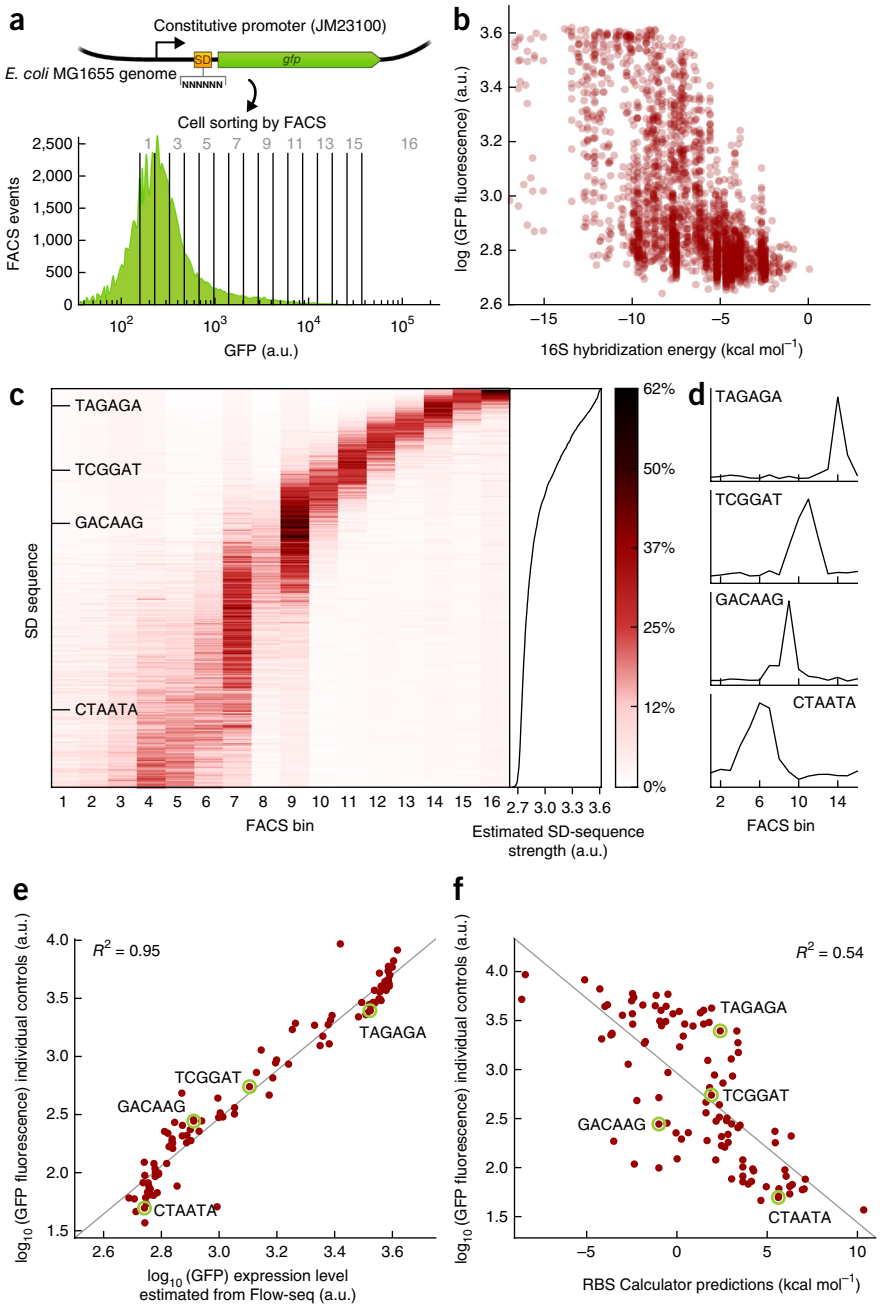

Figure 1 | Characterization of the *E. coli* SD sequence. (**a**) Experimental design and FACS bins with SD-sequence read counts. a.u., arbitrary units. (**b**) Hybridization energies for 16S rRNA and the estimated SD-sequence strength derived from Flow-seq data (n = 3,087). (**c**) Heat map of normalized read counts for SD sequences ordered by estimated SD-sequence strength (left) and estimated SD-sequence strength for all assessed sequences (right) (n = 3,087). Red shading indicates the percentage of total reads (across all bins) for an individual SD sequence that are found in a specific bin. (**d**) Examples of the read counts for four SD sequences across flow-sorted bins. (**e**) Plate reader fluorescence values compared with Flow-seq measurements (n = 106). (**f**) RBS Calculator[15] (version 2.0) predictions compared with observed values for GFP expression (n = 106). Green circles in **e** and **f** indicate sequences shown in **d**.

(version 1.0) to predict the translation-initiation rates for all 4,096 SD sequences and compared those values to GFP levels estimated by Flow-seq (**Supplementary Fig. 1**). We also compared RBS Calculator (version 2.0) predictions for the 106 single-clone GFP levels to the plate reader data (**Fig. 1f**). Both sets of predictions were not very strongly correlated ($R^2 = 0.44$, $P < 10^{-13}$ and $R^2 = 0.54$, $P < 10^{-17}$, respectively), highlighting the need for further development of predictive computational models.

To address the need for better *de novo* models, we developed EMOPEC (freely available at http://emopec.biosustain.dtu.dk and as **Supplementary Software**) to predict *E. coli* gene expression on the basis of SD-sequence engineering. Out of 4,096 possible SD sequences, we experimentally characterized the strength of 3,087, and we predicted the strength of the remaining sequences using a Random Forest regressor (Online Methods). Fivefold cross-validation of the model yielded an R^2 of 0.89 (**Supplementary Fig. 2**). We also developed a method for identifying the SD sequence of a gene, which involves scanning the region upstream of the start codon for the 6-bp sequence with the highest predicted expression and multiplying by a penalty function that compensates for a suboptimal distance to the start codon (Online Methods).

With EMOPEC, identified SD sequences are assigned a strength value on the basis of a model derived from Flow-seq data (**Fig. 1c** and Online Methods). The lowest and highest predicted strength values for a given SD sequence are set to 0% and 100%, respectively, and EMOPEC selects sequences with expression close to the desired target that result in a minimum change in secondary structure. The default output is ten SD sequences predicted to produce ten linearly spaced expression levels from 10% of the maximum up to the predicted maximum expression level (this refers to our estimate of the highest expression that can be achieved by varying the SD sequence alone).

Control elements such as ribosome-binding sites and promoters are frequently context dependent, and expression levels may depend on the N-terminal protein sequence of the expressed gene, especially owing to the effects of the mRNA secondary structure[1–3,15,18]. To minimize these effects, EMOPEC calculates the free energy of secondary structures −35 to +35 nt from the start codon (excluding the contribution from SD-ribosome binding[18]) for all SD sequences substituted into a given transcript. For each targeted relative expression level, EMOPEC samples transcripts with at least ten SD sequences predicted to result in expression closest to the desired value and then filters according to the least predicted change in secondary-structure energy compared with the original transcript sequence. The extent of changes in secondary

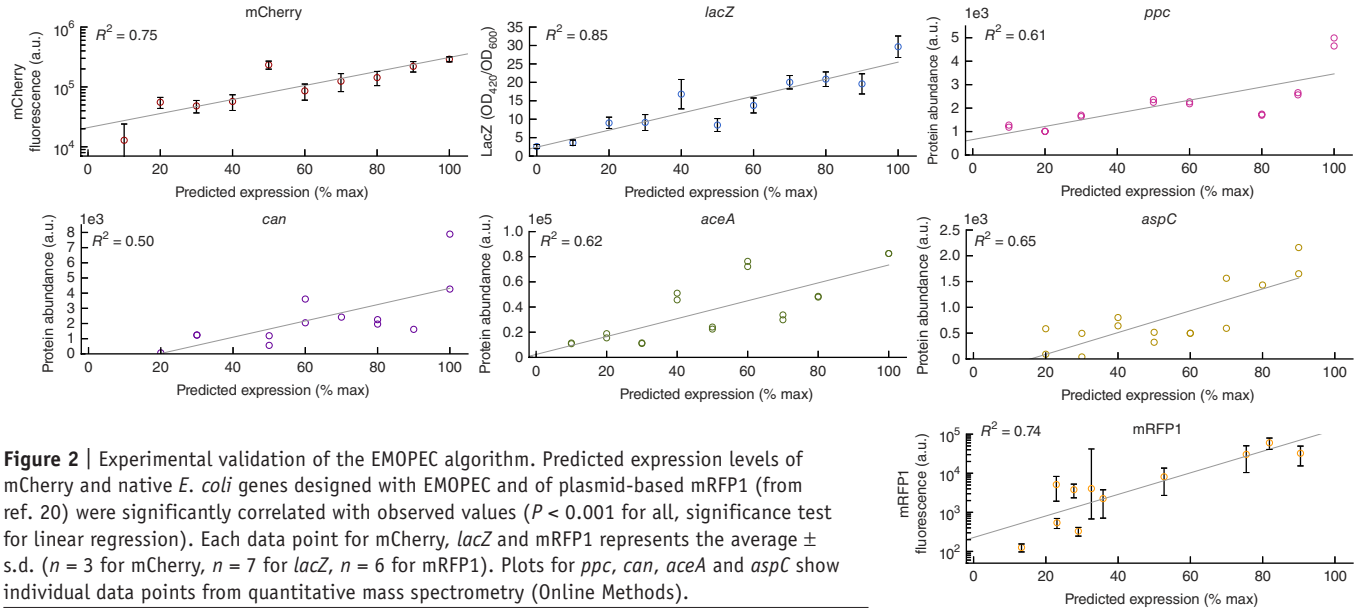

Figure 2 | Experimental validation of the EMOPEC algorithm. Predicted expression levels of mCherry and native *E. coli* genes designed with EMOPEC and of plasmid-based mRFP1 (from ref. 20) were significantly correlated with observed values ($P < 0.001$ for all, significance test for linear regression). Each data point for mCherry, *lacZ* and mRFP1 represents the average ± s.d. ($n = 3$ for mCherry, $n = 7$ for *lacZ*, $n = 6$ for mRFP1). Plots for *ppc*, *can*, *aceA* and *aspC* show individual data points from quantitative mass spectrometry (Online Methods).

structure is predicted to be very low; thus most changes in expression are expected to result from differences in ribosome-binding strength. EMOPEC uses the MAGE Oligo Design Tool[19] (MODEST) to generate optimized MAGE oligos for directly engineering desired protein-expression changes.

We generated a resource that includes ten MAGE oligos predicted to result in linearly spaced expression levels for every gene in *E. coli* K12 MG1655 (**Supplementary Table 3**). One version is an unconstrained library, and a second version introduces synonymous substitutions only into genes that overlap the SD sequence. The change in free energy corresponding to all 40,526 designed SD-sequence mutations in the resource is 0.51 kcal mol^{-1} on average (**Supplementary Fig. 3** and **Supplementary Table 3**), and most SD-sequence substitutions introduce only 3–5-bp changes. Even with the constraints, it was possible to design changes in SD sequence that led to expression levels very close to target values for most genes (**Supplementary Fig. 4**). Changes in SD sequence are slight and are not expected to perturb other aspects of regulation; however, in some cases, complex regulation (for example, involving transcription factor–binding sites, Rho-independent termination sites or riboswitches) may be affected, and users are advised to check for this manually.

To test performance, we used EMOPEC to generate up to ten evenly spaced protein-expression levels for each of six chromosomal genes: integrated *mCherry* and the native *E. coli* genes *lacZ*, *aceA*, *can*, *ppc* and *aspC*. λ-Red recombination was used to introduce the MAGE oligo sequences, and expression was measured on the basis of mCherry fluorescence, β-galactosidase activity for *lacZ*, or quantitative mass spectrometry for the remaining genes (Online Methods). We observed a linear relationship and good correlation between EMOPEC-predicted and observed expression for all tested proteins (**Fig. 2**; $R^2 = 0.55–0.85$). We also tested the algorithm on published data from a plasmid system in which an SD sequence was varied and downstream mRFP1 was measured[20]. The results indicated that EMOPEC can also be used to design relative expression levels in a plasmid system (**Fig. 2**; $R^2 = 0.78$). We did not observe significant changes in *mCherry* mRNA levels among the eight strains tested, indicating that differences in expression were not due to changes in transcriptional activity (**Supplementary Fig. 5** and Online Methods).

Even though we anticipated that the detection of native ribosome-binding sites might not be precise in all cases, the results for the six tested genes show that predictions were sufficiently precise for EMOPEC to function. We recommend that the option to manually specify the spacing be used whenever possible.

We also compared observed expression levels for the six test genes with predictions from RBS Calculator version 2.0 (**Supplementary Figs. 6** and **7**). The results showed that correlations were lower with RBS Calculator (pooled $R^2 = 0.37$, compared with $R^2 = 0.64$ for EMOPEC).

Our data suggest that it is possible to predictably modulate relative protein-expression levels through carefully designed changes in upstream SD sequences. Our engineering design strategy is based on empirical data and facilitates the uniform exploration of phenotypic space in *E. coli*. EMOPEC reduces the number of mutations that need to be constructed for metabolic engineering through the use of small changes that can be implemented with high efficiency and without selection markers. As an example, we constructed more than 60 individual strains to test EMOPEC within 3 weeks, from design to sequence validation of single colonies.

With EMOPEC, 91% of the designed sequences led to measured protein levels within twofold of the desired target level. In comparison, state-of-the-art studies with standardized genetic elements showed reliability of 64% (ref. 3) and 93% (ref. 2), and RBS Calculator was reported to have a 47% probability of expressing a protein to within twofold of the target level[15]. Furthermore, when modifying genomically encoded genes or pathways, especially in multiplex, insertion of standardized genetic elements is often not possible, whereas EMOPEC can be used to change the expression level of both chromosomal and plasmid-encoded genes.

Finally, the data sets used to build these data-driven models can also be used to refine and parameterize *de novo* models, thereby

increasing the understanding of the biophysical principles governing the control of transcription and translation.

METHODS

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

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

ACKNOWLEDGMENTS

We thank H. Genee, A. Wallin, S. Cardinale and H. Wang for discussions and suggestions regarding this manuscript, and we thank A. Koza for assistance with DNA sequencing. The research leading to these results received funding from the Novo Nordisk Foundation through the Novo Nordisk Foundation Center for Biosustainability and the European Union Seventh Framework Programme (FP7-KBBE-2013-7-single-stage) under grant agreement 613745, Promys.

AUTHOR CONTRIBUTIONS

M.T.B., M.P., M.S.K., S.I.J., T.W. and S.H. conducted the experiments. M.T.B., M.S.K. and M.J.H. conducted bioinformatics and data analysis. A.T.N. supervised the flow cytometry experiments. S.H. supervised the proteomics experiments. M.O.A.S., M.T.B., M.P. and M.S.K. designed the study. M.O.A.S. conceived and supervised the project. M.T.B., M.S.K. and M.O.A.S. wrote the manuscript, and all authors contributed to editing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mutalik, V.K. *et al. Nat. Methods* **10**, 347–353 (2013).
2. Mutalik, V.K. *et al. Nat. Methods* **10**, 354–360 (2013).
3. Kosuri, S. *et al. Proc. Natl. Acad. Sci. USA* **110**, 14024–14029 (2013).
4. Goodman, D.B., Church, G.M. & Kosuri, S. *Science* **342**, 475–479 (2013).
5. Lee, J.W. *et al. Nat. Chem. Biol.* **8**, 536–546 (2012).
6. Woolston, B.M., Edgar, S. & Stephanopoulos, G. *Annu. Rev. Chem. Biomol. Eng.* **4**, 259–288 (2013).
7. Wang, H.H. *et al. Nature* **460**, 894–898 (2009).
8. Bonde, M.T. *et al. ACS Synth. Biol.* **4**, 17–22 (2015).
9. Sommer, M.O., Church, G.M. & Dantas, G. *Mol. Syst. Biol.* **6**, 360 (2010).
10. Klumpp, S., Zhang, Z. & Hwa, T. *Cell* **139**, 1366–1375 (2009).
11. Gold, L. *Annu. Rev. Biochem.* **57**, 199–233 (1988).
12. Shine, J. & Dalgarno, L. *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346 (1974).
13. Schurr, T., Nadir, E. & Margalit, H. *Nucleic Acids Res.* **21**, 4019–4023 (1993).
14. Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E. & Schneider, T.D. *J. Mol. Biol.* **313**, 215–228 (2001).
15. Salis, H.M. *Methods Enzymol.* **498**, 19–42 (2011).
16. Reeve, B., Hargest, T., Gilbert, C. & Ellis, T. *Front. Bioeng. Biotechnol.* **2**, 1–6 (2014).
17. Seo, S.W. *et al. Metab. Eng.* **15**, 67–74 (2013).
18. Salis, H.M., Mirsky, E.A. & Voigt, C.A. *Nat. Biotechnol.* **27**, 946–950 (2009).
19. Bonde, M.T. *et al. Nucleic Acids Res.* **42**, W408–W415 (2014).
20. Farasat, *et al. Mol. Syst. Biol.* **10**, 731 (2014).

ONLINE METHODS

Bacterial strains, plasmids and reagents. Strains and plasmids used for recombineering, fluorescence measurements and β-galactosidase enzymatic-activity assays in this study are listed in **Supplementary Table 4**. All strains were grown in lysogeny broth (LB; 10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl) or on LB-agar plates supplemented with appropriate antibiotics (ampicillin, 100 μg/ml; kanamycin, 50 μg/ml) if needed, or grown in M9 minimal media (6.8 g/L Na_2PO_4, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$) supplemented with trace elements (0.5 mg/L $FeCl_3{\cdot}6H_2O$, 0.09 mg/L $ZnSO_4{\cdot}7H_2O$, 0.06 mg/L $CuCl_2{\cdot}2H_2O$, 0.06 mg/L $MnSO_4{\cdot}H_2O$, 0.09 mg/L $CoCl_2{\cdot}6H_2O$), Wolfe's vitamin solution (10 μg/L pyridoxine hydrochloride, 5 μg/L thiamine-HCl, 5 μg/L riboflavin, 5 μg/L nicotinic acid, 5 μg/L calcium D-(+)-pantothenate, 5 μg/L p-aminobenzoic acid, 5 μg/L thiotic acid, 2 μg/L biotin, 2 μg/L folic acid, 0.1 μg/L vitamin B12), and 2 g/L glucose. Minimal M9 agar plates were supplemented with 1.5% agar. All oligos were synthesized by Integrated DNA Technologies (Leuven, Belgium) and are listed in **Supplementary Table 4**. PCR reactions were performed using Phusion High-Fidelity DNA polymerase (New England BioLabs) according to the manufacturer's instructions.

Construction of fluorescent *E. coli* strains and pMA1. *gfp* and *mCherry*[21] were PCR amplified with oligos designed to contain the constitutive promoter sequence BBa_J23100 (Registry of Standard Biological Parts) in combination with a strong SD sequence using, respectively, a folding reporter vector and pKS1 (refs. 22,23) (obtained from Morten Nørholm, DTU) as templates. An FRT-flanked kanamycin cassette was PCR amplified from pkd4 (ref. 24), and they were spliced together with the PCR-amplified fluorescent genes by overlap extension PCR. The spliced cassettes were integrated into the genome of *E. coli* K12 MG1655 (9 bp downstream of *glmS*) by λ-red recombineering using the temperature-sensitive plasmid pkd46 (ref. 24). The FRT-flanked KmR cassettes were subsequently removed using the temperature-sensitive, FLP-containing plasmid pCP20 (ref. 25). The GFP variant originally isolated from *Aequorea victoria* contains the following substitutions compared to the sequence reported by Prasher *et al.*[26]: F64L, S65T, Q80R, F99S, M153T, V163A and I219V. Cells constitutively expressing *gfp* were detectable under blue light.

Plasmid pMA1 was constructed by PCR amplification of the λ-phage β-gene from strain EcNR2 (ref. 7). The PCR fragment was digested with NcoI and HindII and cloned into the corresponding sites of pBAD24. The insert was verified by sequencing (Beckman Coulter Genomics, UK).

Single-stranded oligo recombineering. Cells containing the pMA1 plasmid were grown in 15 ml of LB broth supplemented with ampicillin (100 μg/ml) with shaking at 37 °C to an OD_{600} of 0.4, after which the β-protein was induced for 10 min by the addition of arabinose to a final concentration of 0.2%. After induction, cells were placed on ice for 15 min before they were harvested, washed, and finally resuspended in a total volume of 200 μl of ice-cold sterile water. We mixed 50 μl of electrocompetent cells with 5 pmol of oligo and electroporated them in 0.1-mm gap cuvettes at 1.8 kV, 200 Ω and 25 μF. Immediately after electroporation, we added 1 ml of LB to the cells. Cells were transferred to a 50-ml Falcon tube to a total volume of 5 ml in LB and grown for at least 3 h at 37 °C to allow segregation of chromosomal DNA. After outgrowth, the culture was diluted and spread on LB-agar plates for further analysis.

SD-*gfp* cell library. We used single-stranded (ss) oligo recombineering to change the 6-nt SD sequence (AGGAGA) located 8 bases upstream of the *gfp* initiation codon by using 90-bp-long ssDNA oligos. Initially the SD sequence was altered to the anti-SD sequence (TCCTCC) using oligo P324. The recombineering efficiency was approximately 10%. We isolated single colonies by re-streaking at least twice, and we verified the insert by sequencing. A single colony containing the anti-SD sequence was inoculated and used to create an SD-*gfp* cell library using oligo P329 that included a 6-bp randomized sequence (N6) covering the SD sequence. Six consecutive rounds of recombineering were performed, resulting in <1% fluorescent colonies. In a parallel recombineering experiment using oligo P325, which contained the consensus SD sequence (AGGAGG), 28% of the cells were fluorescent after six rounds of recombineering. This suggests that the SD-*gfp* cell library contained about 30% recombinant cells, with only a fraction of the introduced sequences (<1%) resulting in visual GFP levels.

Cell sorting of SD-*gfp* library using flow cytometry. We analyzed an exponentially growing culture containing the SD-*gfp* cell library by flow cytometry in order to bin cells according to their level of fluorescence. Cell sorting was done in precision single-cell mode with an event rate of around 3,000 per second. Sorted cells were collected in microtiter plates containing 100 μl of LB. For cell sorting, 16 gates were made so that cells with similar fluorescence levels were collected in the same bin. Sorting resulted in 50,000 events in gates P6–P11, 20,000 in gates P12–P15, 5,000 in gates P16–P20 and 1,000 in gate P21. The collected cells were grown overnight in 1.5 ml of LB with shaking at 37 °C. We stored 0.5 ml of cell culture from each library as glycerol stock at −80 °C, and we harvested 0.6 ml of cell culture for extraction of chromosomal DNA.

Next-generation sequencing of SD sequence–containing amplicons. We extracted chromosomal DNA using the DNeasy Blood and Tissue kit (Qiagen) and eluted it in 200 μl of elution buffer. We subsequently used 1 μl of DNA as a template for PCR amplification. For each of the individually sorted pools, we used a specific barcoded primer set (**Supplementary Table 5**) to amplify a 103-bp DNA fragment containing the SD-sequence region of interest. We purified PCR products using the DNA Clean and Concentration kit (Zymogen Research, USA). We quantified the PCR fragments using a Qubit 2.0 Fluorometer (Life Technologies) and pooled them in equal quantities to a total amount of 1 μg. We barcoded the PCR pool using Illumina TruSeq Adaptor Index 1, amplified it according to the TruSeq manual, and sequenced it on an Illumina MiSeq sequencer using 150-bp paired-end reads (SciLifeLab, Karolinska Institutet Science Park, Sweden).

Determination of fluorescence levels of individual SD-*gfp* strains. From each of the 16 sorted bins, we isolated eight individual colonies by re-streaking twice on LB-agar plates.

The SD-sequence region of interest was sequenced, and a total of 106 individual strains were identified. The SD sequences of the individual isolates are shown in **Supplementary Table 2**.

Single colonies grown on M9 minimal plates were inoculated in 150 µl of M9 minimal media supplemented with 0.2% glucose and grown overnight at 37 °C with shaking in 96-well microtiter plates in a BioTek ELx808 absorbance microplate reader. We transferred 4 µl of the overnight cell cultures to 146 µl of M9 media and incubated them in 96-well microtiter plates with shaking at 37 °C until the cultures reached an OD_{450} between 0.3 and 0.4. Finally, we transferred 6 µl of each culture to a flat-bottom microtiter plate containing 200 µl of 1× PBS for flow cytometry measurement by FACS Fortessa.

Flow-seq data analysis. We used Qiime (http://qiime.org/) to demultiplex the reads using the custom sequence barcodes (**Supplementary Table 5**) at either 5′ or 3′ ends of the reads, and to extract the SD sequences (split_libraries_fastq.py, split_libraries.py and andjust_seq_orientation.py scripts in Qiime). The following options were used for split_libraries.py: allow five mismatches in the regions flanking the SD sequence; minimum read length, 50; maximum read length, 200. We determined the numbers of distinct SD sequences in each demultiplexed flow-sorted bin using a custom Python script that is available upon request. A total of 3,880 distinct SD sequences were detected at least once in the whole data set. The SD sequence counts in each flow-sorted bin are available in **Supplementary Table 6**.

We used the SD sequence–count table to derive estimates of GFP expression for each SD sequence using custom Matlab scripts. First, a single Gaussian curve was fit to the major peaks in the signal-intensity curves for each flow-sorted bin to obtain the mean (e_b) and variance of the GFP expression for each bin b (**Supplementary Fig. 8**). Data for gate P5 were excluded from further analysis, as the estimated mean expression level for this gate was higher than that for the next gate, P6. To estimate the level of GFP expression for each SD sequence, we applied the method first introduced by Sharon *et al.*[27]. This method computes the weighted mean expression level of SD sequence s (f_s) using the formula $f_s = (\Sigma_b\, n_{b,s}/n_b \times e_b)/(\Sigma_b\, n_{b,s}/n_b)$, where e_b is the mean value of the Gaussian fit mentioned above, n_b is the total count of all SD sequences in bin b, and $n_{b,s}$ is the count of SD sequence s in bin b.

The distribution of total merged read counts across all bins for each SD sequence was bimodal (**Supplementary Fig. 9**) with a peak at ~1,000 reads and a second peak at one read. To obtain high-confidence estimates of GFP expression levels from Flow-seq, we included only SD sequences with at least 50 merged reads across all flow-sorted bins. The final high-confidence data set included 3,087 SD sequences (**Fig. 1c**). The full list of GFP expression levels for each SD sequence estimated using Flow-seq is provided as **Supplementary Table 1**. **Supplementary Table 2** contains a list of the 106 single-clone expression levels measured directly, estimated using Flow-seq and computed using RBS Calculator version 2.0 (**Fig. 1e**,**f**).

Calculation of RBS Calculator–predicted translation-initiation rates of SD sequences. In order to compare the experimental expression levels with levels obtained using a state-of-the-art computational model, we used RBS Calculator[15]. We calculated the expected translation-initiation rate of each of the SD sequences using an online version of the calculator (version 2.0) in reverse-engineering mode.

Design of algorithm to modify gene expression levels. The EMOPEC algorithm is published online at http://emopec.biosustain.dtu.dk, and the source code with comments is available in the **Supplementary Software**. The web server input is two sequences divided into the coding sequence and the 5′ untranslated region, where the ribosome-binding site is located. Optionally, a sequence containing design constraints may be supplied in IUPAC degenerate base notation format. Users can adjust the prediction of the SD sequence either by choosing to let EMOPEC predict the location of the SD sequence or by directly supplying the spacing distance given as the number of nucleotides between the SD sequence and the initiation codon.

In the case of automatic prediction, EMOPEC maximizes the predicted expression value by looking at all potential SD sequences with spacing in the range of 1–13 nt. The predicted expression, which is maximized, is given by the following equation:

$$\text{Expression(SD,Spacing)} = \text{EMOPEC(SD)} - \Delta G_{\text{spacing}} \times c$$

where EMOPEC(SD) is a table lookup in the EMOPEC library (in log(GFP)), $\Delta G_{\text{spacing}}$ is a spacing penalty[18], and c is a constant used to scale the penalty to the arbitrary units used by EMOPEC. We estimated c as 0.235 by maximizing the recovery rate of values from the EMOPEC library when applied to the original GFP sequence. We accomplished this by scanning a large number of c values and choosing the value that led to correct prediction of the largest number of SD sequences.

The predicted expression and library expression values are given as a percentage of the maximum possible expression value, calculated with the equation

$$\text{Expression}_{\text{percent}}\text{(SD,Spacing)} = \frac{\text{Expression(SD,Spacing)} - \text{Expression(TTGGGC,Spacing)}}{\text{Expression(AGGAGA,Spacing)} - \text{Expression(TTGGGC,Spacing)}}$$

where TTGGGC is the lowest expression level in the EMOPEC library and AGGAGA is the highest expression level. Finally, oligos are created using MODEST[19], with the new SD sequences as input and the original sequence as the reference sequence.

The library is created between two setpoints, which the user chooses by selecting one of three options: "up," which will start the library at the predicted expression level and end it at the highest possible expression level; "down," which will start at the lowest possible expression level and end at the predicted expression level; and "both," which will ignore the predicted expression level, starting the library at the lowest possible expression level and ending it at the highest possible expression level.

The tool calculates the library by first creating a list of linearly spaced target expression levels on the basis of the given options. For instance, if "both" is chosen and the library size is set to 4, relative levels of 100%, 75%, 50% and 25% are set as target expression values. For each target expression level, ten candidate new SD sequences with expression levels centered on the desired new target are sampled from the EMOPEC library. For example, given 50% target expression, SD sequences from TTTGGA with predicted relative expression of 49.59% to TGCGGT with 50.25%

doi:10.1038/nmeth.3727

relative predicted expression and eight additional sequences in between are selected as candidate SD sequences. It then calculates the secondary structure energy delta ($\Delta\Delta G$) by estimating the folding minimum free energy of the original and mutated full sequence using RNAfold from the ViennaRNA package[28]. The SD sequence with the lowest $\Delta\Delta G$ is chosen as the target sequence.

If a sequence with constraints is supplied, only the subset of sequences in the EMOPEC library satisfying the constraints are considered. This may lead to unevenly spaced or smaller libraries than expected, and the user is responsible for the final evaluation of the library.

The algorithm is implemented in Python 2.7 using the SWIG (http://www.swig.org/) interface to the ViennaRNA package[28] (http://www.tbi.univie.ac.at/RNA/). The server front-end is implemented using the AngularJS framework (https://angularjs.org/), sending JSON requests to a back-end server implemented in Python 2.7 using the Flask framework (http://flask.pocoo.org/).

We used EMOPEC to design ten MAGE oligos for all genes in *E. coli* K12 MG1655, predicted to result in linearly spaced expression levels from a low level to the maximum predicted level possible by changing the SD sequence. This led to a resource (available at http://emopec.biosustain.dtu.dk/optilib and in **Supplementary Table 3**) comprising 40,526 different MAGE oligos, which enabled exploration of the protein expression levels of all currently predicted protein-coding genes in *E. coli* K12 MG1655.

The SD-sequence location in the input sequence is either specified directly by the user or found by a search of the leading region for the SD sequence with the highest expression level multiplied by a previously developed penalty function. The library is created by first sampling new SD sequences in a linear stepwise manner and picking the sequence with the lowest secondary structure energy delta compared to the original sequence. The initial sampling is done by calculating the exact values to create a linear library and choosing sequences with a maximum deviation from the ideal value. The secondary structure energy delta ($\Delta\Delta G$) is calculated by folding the original and mutated full sequence using RNAfold from the ViennaRNA package[28]. Oligos are created using MODEST[19], with the new SD sequences as input and the original sequence as the reference sequence.

To assign a value to the remaining SD sequences, we used a Random Forest regressor. Each sequence in the EMOPEC data set was encoded into a vector containing either 0 or 1 in a one-encoding scheme. Each nucleotide was assigned a vector consisting of three 0's and a single 1—that is, $C = [1, 0, 0, 0]$, $G = [0, 1, 0, 0]$, $T = [0, 0, 1, 0]$ and $A = [0, 0, 0, 1]$. An SD sequence was encoded by concatenating six vectors corresponding to the six nucleotides in the SD sequence, resulting in a final feature vector of length 24. The Random Forest implementation of the Python package scikit-learn (http://scikit-learn.org/) was used to train a Random Forest regressor with 100 trees.

To validate the model, we used fivefold cross-validation as well as the out-of-bag (OOB) score. We calculated the OOB score by predicting all samples individually on a subset of trees in the Random Forest in which the particular sample was not used in the training. The OOB score was calculated to an R^2 value of 0.90. We carried out fivefold cross-validation by splitting the EMOPEC data set into five equal random subsets. For each subset, the other four subsets were used to train a new and independent Random Forest. The sequences in the subset not used in the training were then predicted using the new model, and the process was repeated five times, once for each subset. An R^2 value of 0.89 was calculated using the cross-validation approach (**Supplementary Fig. 2**).

Validation of EMOPEC. To test the functionality and performance of EMOPEC, we used the algorithm to modify the expression level of six additional genes: constitutively expressed *mCherry* integrated in the genome of *E. coli* K12 MG1655, and the native *E. coli* genes on the chromosome *lacZ*, *aceA*, *can*, *ppc* and *aspC*. EMOPEC was used to design ten different MAGE oligos that were predicted to result in ten different, evenly distributed protein levels. Seven out of 60 strains were not constructed or measured, because no mutants were isolated after the MAGE cycles, and we concluded that 53 strains and >8 EMOPEC designs for each gene were sufficient. λ-Red recombineering using pMA1 was performed as described above to introduce the sequences. We subsequently quantified mCherry expression by fluorescence measurement and LacZ expression by measuring β-galactosidase enzymatic activity according to the methods described by Griffith and Wolf[29]. We measured amounts of AceA, Can, Ppc and AspC using the proteomics methods described below. EMOPEC was used with an external, previously published data set in which a plasmid system with an SD sequence is varied, and expression of downstream mRFP1 was measured[20]. The measurements of the original data set were processed with the EMOPEC algorithm, and the filtered sequences with a secondary structure of <2 kcal were plotted, showing $R^2 = 0.78$. For comparison, the secondary structure is on average 0.52 kcal for most of the designed sequences for all *E. coli* genes (**Supplementary Fig. 3**). To calculate the percentage of designed sequences that had measured protein levels within twofold of the desired target level, we divided the value of the point on the linear regression fit corresponding to the predicted expression by the measured values, for all pairs of predicted and measured values. If the resulting value was between 0.5 and 2, the measurement was counted as within twofold. We also predicted the expression levels of the constructs using RBS Calculator[15] to compare predicted versus observed expression levels between EMOPEC and RBS calculator (**Supplementary Fig. 7**).

Preparation of *E. coli* cells for proteomics. Frozen cells were kept at −80 °C for up to 4 weeks, after which they were thawed on ice and pelleted by centrifugation at 15,000*g* for 10 min. The supernatant was removed, and 100 µl of urea (8 M, 75 mM NaCl, 50 mM Tris-HCl, pH 8.2) was added to the samples together with two 3-mm zirconium oxide beads (Glen Mills, NJ, USA). Cells were disrupted using a Mixer Mill (MM 400 Retsch, Haan, Germany) for 2 min at 25 Hz. The samples were then kept at 4 °C for 30 min followed by 2 min at 25 Hz in the Mixer Mill. An additional 100 µl of urea was added, after which samples were subjected to 2 min at 25 Hz in the Mixer Mill and then left for 30 min at 4 °C and a final 2 min at 25 Hz in the Mixer Mill. Samples were centrifuged at 15,000*g* for 10 min, and 100 µl of supernatant were collected and diluted with 400 µl of 25 mM ammonium bicarbonate, after which the volume was reduced to 100 µl using a 3 kDa–cutoff filter. Protein concentrations were measured, and 100 µg were used for tryptic digestion. Prior to digestion, 5 µl of 100 mM DTT was added and samples were kept at 37 °C for 45 min, after which 10 µl of 100 mM iodoacetamide was added and samples

were kept in the dark for 45 min. Tryptic digestion was carried out for 8 h, after which 10 µl of 10% TFA was added and samples were StageTipped using C18 (Empore, 3M, USA) according to the procedure described by Rappsilber *et al.*[30].

Nanoscale LC separation of the tryptic digested samples, each containing 1 µg of protein, was performed using a nanoACQUITY system (Waters, USA) equipped with a Symmetry C18 5-µm, 180 µm × 20 mm precolumn and a nanoACQUITY BEH130 C18 1.7-µm, 75 µm × 250 mm analytical reversed-phase column (Waters, USA). Initially the samples were trapped on the precolumn using mobile phase A, consisting of 0.1% formic acid in water with a flow rate of 8 µL min^{-1} for 4 min. Mobile phase B consisted of 0.1% formic acid in acetonitrile. A reversed-phase gradient was used to separate peptides going from 5% to 40% acetonitrile in water over 90 min with a flow rate of 250 nL min^{-1} and a constant column temperature of 35 °C.

Eluates were immediately sprayed into a Synapt G2 (Waters, Manchester, UK) Q-ToF instrument operated in positive mode using electrospray ionization with a NanoLock-spray source. Leucine encephalin was used as a lock mass, supplied from the internal fluidics system of the mass spectrometer. The lock mass channel was sampled every 60 s. For all samples, the mass spectrometer was operated in resolution mode, with continuum spectra being acquired. The mass spectrometer alternated between low- and high-energy modes using a scan time of 0.8 s for each mode over 50–2,000 Da. In the low-energy MS mode, data were collected at a constant collision energy of 4 eV. In the elevated-energy MS mode, the collision energy was increased from 15 to 40 eV.

Proteomics data analysis. We obtained protein identification and quantification data by using Progenesis QI for Proteomics version 2.0 and the *E. coli* K-12 MG1655 UniProt proteome database (ID UP000000625). Only unique peptides of the proteins of interest were used for quantification, enabling comparisons of protein abundance across the different samples[31]. Pre-established conditions for exclusion of samples were a protein level below the detection threshold for >75% of the samples.

RNA extraction. We transferred 5 µl of overnight cultures to 5 ml of LB-ampicillin and allowed them to grow in 50-ml Falcon tubes to an OD_{600} of 0.4–0.6 (mCherry); the cultures were then vortexed and placed on ice for 5 min before being centrifuged for 2 min at 6,500*g*, after which cell pellets were frozen in liquid nitrogen and stored at −80 °C until RNA extraction. Cells were digested for 5 min at 25 °C in TE buffer containing 1 mg ml^{-1} lysosome and 1 mg ml^{-1} proteinase K, after which total RNA was extracted using the RNeasy mini kit (Qiagen Sciences, Maryland, USA) according to the manufacturer's instructions. Extracted RNA was digested with DNase I, and extraction was followed by a phenol:chloroform extraction and ethanol precipitation, after which PCRs were performed on non–reverse transcripts to verify the removal of DNA. The integrity of the RNA was verified on 1% agarose gels, and purity and quantification of total RNA were assessed on a Nanodrop 2000 (Nanodrop Thermo Scientific, USA).

Reverse-transcription quantitative PCR. Random primers were annealed with 0.5 µg of total RNA using the SuperScript III First-Strand synthesis system for RT-PCR (Invitrogen Corp., Carlsbad, CA, USA) according to the manufacturer's instructions. Synthesized cDNA was diluted tenfold, and 2 µl of diluted cDNA was used as a template for quantitative PCR using SYBR GreenER qPCR SuperMix Universal (Invitrogen Corp., Carlsbad, CA, USA) and the Mx3000P qPCR system (Agilent Technologies, Santa Clara, California, USA) according to the manufacturer's instructions. Specific amplification of target genes was 1 cycle at 50 °C for 2 min, 1 cycle at 95 °C for 10 min, and 40 cycles at 95 °C for 15 s, 60 °C for 1 min, and 82 °C for 10–15 s. The 82 °C step was introduced to minimize the effect of primer dimers, which was observed during melting-curve analysis for the reference gene *frr*. The relative expression levels of individual transcripts were determined according to the method described by Pfaffl[32], using *frr* as a reference gene, as has been done previously[33].

Code availability. The source code for EMOPEC, including the web server, is available as **Supplementary Software**.

Reproducibility. Sample sizes for each experiment were chosen on the basis of initial pilot experiments and similar experiments in the literature. No blinding or randomization was used in the experiments conducted. Data from bin P5 in flow cytometry experiments were excluded from further analysis because the estimated mean expression level for this gate was higher than that for the next gate, P6.

21. Shaner, N.C. *et al. Nat. Biotechnol.* **22**, 1567–1572 (2004).
22. Waldo, G.S., Standish, B.M., Berendzen, J & Terwilliger, T.C. *Nat. Biotechnol.* **17**, 691–695 (1999).
23. Söderström, B.*et al. Mol. Microbiol.* **92**, 1–9 (2014).
24. Datsenko, K.A. & Wanner, B.L. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6640–6645 (2000).
25. Cherepanov, P.P. & Wackernagel, W. *Gene* **158**, 9–14 (1995).
26. Prasher, D.C., Eckenrode, V.K., Ward, W.W., Prendergast, F.G. & Cormier, M.J. *Gene* **111**, 229–233 (1992).
27. Sharon, E. *et al. Nat. Biotechnol.* **30**, 521–530 (2012).
28. Lorenz, R. *et al. Algorithms Mol. Biol.* **6**, 26 (2011).
29. Griffith, K.L. & Wolf, R.E. *Biochem. Biophys. Res. Commun* **290**, 397–402 (2002).
30. Rappsilber, J., Mann, M. & Ishihama, Y. *Nat. Protoc.* **2**, 1896–1906 (2007).
31. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).
32. Pfaffl, M.W. *Nucleic Acids Res.* **29**, e45 (2001).
33. Herring, C.D. & Blattner, F.R. *J. Bacteriol.* **186**, 6714–6720 (2004).

doi:10.1038/nmeth.3727