Quantitative estimation of activity and quality for collections of functional genetic elements

Vivek K Mutalik, Joao Guimarães, Guillaume Cambray, Quynh-Anh Mai, Marc Juul Christoffersen, Marc Juul Christoffersen, Lance Martin, Ayumi Yu, Colin Lam, César Rodríguez, et al.

Quantitative estimation of activity and quality for collections of functional genetic elements

Vivek K Mutalik[1–3,9], Joao C Guimaraes[1,3,4,9], Guillaume Cambray[1,3,9], Quynh-Anh Mai[1,3], Marc Juul Christoffersen[1,3], Lance Martin[1,3,8], Ayumi Yu[1,3,8], Colin Lam[1,3], Cesar Rodriguez[1,3,8], Gaymon Bennett[1,3,8], Jay D Keasling[1–3,5,6], Drew Endy[1,7,9] & Adam P Arkin[1–3,9]

The practice of engineering biology now depends on the *ad hoc* reuse of genetic elements whose precise activities vary across changing contexts. Methods are lacking for researchers to affordably coordinate the quantification and analysis of part performance across varied environments, as needed to identify, evaluate and improve problematic part types. We developed an easy-to-use analysis of variance (ANOVA) framework for quantifying the performance of genetic elements. For proof of concept, we assembled and analyzed combinations of prokaryotic transcription and translation initiation elements in *Escherichia coli*. We determined how estimation of part activity relates to the number of unique element combinations tested, and we show how to estimate expected ensemble-wide part activity from just one or two measurements. We propose a new statistic, biomolecular part 'quality', for tracking quantitative variation in part performance across changing contexts.

Genetic engineers must specify a priori the precise activities of biomolecular parts for use in integrated synthetic systems[1–11]. Improvements in methods and tools for synthesizing and assembling DNA[12,13] additionally challenge practitioners to design genetic sequences that result in precise expression of hundreds of coding sequences[14–16]. Meanwhile, distributed communities of researchers struggle to collectively assemble, measure, validate and share collections of standard biological parts[17,18]. Additionally, sophisticated biotechnology applications addressing medical or environmental needs demand improved tools for reliably estimating the expected performance of engineered systems[19].

Against this backdrop of needs, studies of biological part activities[20–24] have revealed that the quantitative activity of genetically encoded elements is often highly context dependent[15,25–29]. For example, engineers and biologists have generated and studied libraries of synthetic expression control elements[21,30–38] on an *ad hoc* basis and across relatively limited contexts[14,39]. In some cases, researchers have used first-principle models to develop predictors of element function that attempt to account for context impacts[30,36,37,40–42]. Though valuable, these models cannot fully capture the impact of changing contexts on genetic element function. More recently, researchers have developed passive and active genetic insulators such that the functioning of one element might not corrupt a neighboring element[43–45]. Yet, with the lack of systematic and quantitative data detailing how and to what extent different types of genetic elements interact, it remains unclear that such projects have focused on regularizing the most difficult element-element junctions or leveraged the simplest normalizing molecular mechanisms.

We developed an easy-to-deploy mathematical framework that can be used to score the intrinsic activities of genetic elements to track how such activities vary (or not) across changing contexts. We propose that variation in part activity can serve as a quantitative score of part quality to concisely summarize the reliability of part reuse. For example, we might score a 'promoter' element that never initiates transcription in any and all contexts as a 'high-quality promoter encoding zero activity', whereas an element that in some contexts initiates transcription and in others does not, would be a 'low-quality promoter encoding intermediate activity'. Genetic element 'quality' in these two examples captures the extent to which users of elements can rely on the reported behaviors; a promoter that is known to never initiate transcription would be of particular value for establishing negative controls used in quantifying both transcription promoters and terminators.

To develop and demonstrate the method, we constructed a full combinatorial library of frequently used transcription and translation elements in *E. coli*, expressed two genes at two temperatures

[1]BIOFAB International Open Facility Advancing Biotechnology, Emeryville, California, USA. [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, 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]Present addresses: Department of Bioengineering, Stanford University, Stanford, California, USA (L.M.); Philotic, Inc., San Francisco, California, USA (A.Y.); Autodesk, Inc., San Francisco, California, USA (C.R.); and Center for Biological Futures, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA (G.B.). [9]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 24 OCTOBER 2012; ACCEPTED 13 FEBRUARY 2013; PUBLISHED ONLINE 10 MARCH 2013; DOI:10.1038/NMETH.2403

and measured individual amounts of mRNA and protein for all element combinations. We quantified element activity and quality using the full factorial experimental data set. We identified the junction between translation initiation elements and downstream genes as the major source of irregular gene expression and demonstrated that only a few measurements are needed to estimate the activity of new elements to within reasonable accuracy, once an initial combinational landscape has been established.

RESULTS

Quantifying context effects as a score of part quality

We used a linear model of gene expression (based on a conventional model)[46] to track population average steady-state protein levels and to study element-element context effects (see Online Methods for equations). We assumed that expression output signals (such as arbitrary fluorescence levels and transcript abundances) measured for expression cassettes are a function of promoter, 5′ UTR and gene of interest (GOI) elements plus interactions among elements; genetic elements are formally defined below.

Although functional relationships among genetic elements and changing environments will not always map to such simple, molecular mechanism–agnostic linear models, these approximations were appropriate for our underlying goal: to establish a basic framework that many researchers might easily use to contribute to the estimation of activities and quality of biological parts and to share such information to improve part collections. Stated differently, the framework developed above is primarily focused on the recording and reporting of measurements and not on deeper mechanistic understanding. By enabling a collective capacity to identify categories of low-quality parts and problematic element-element junctions, we sought a means to enable, prioritize and evaluate subsequent work to better understand and ultimately engineer higher-quality genetic elements[47].

Experimental design

Many extrinsic factors can overwhelm observed variation in the activities of transcription-control and translation-initiation elements[15,48,49]. Thus, a first challenge was to determine whether we could directly observe subtle or modest quantitative variation in genetic element activities arising from only the reuse of parts in combination. To do so, we first performed carefully controlled replicate experiments under common physical conditions.

We selected widely used, representative genetic elements encoding transcription-control and translation-initiation functions; although hereafter we refer to each category of control elements as ‘promoters’ or ‘5′ UTRs’, we note that our selected elements are encoded by irregular DNA sequences as reported and as typically used elsewhere (**Supplementary Table 1**). For example, promoters may include DNA sequence beyond the transcription start that would contribute promoter-associated mRNA sequence to any coupled 5′ UTR and thereby potentially modulate both translation initiation and mRNA stability. Moreover, the DNA sequence after a transcription start site is also known to modulate RNA polymerase promoter escape and, hence, could also affect promoter strength[21]. The total number of nucleotides preceding translation initiation codons varies from 21 to 59 across the mRNA encoded by the promoters and 5′ UTRs assembled here (**Supplementary Table 1**).

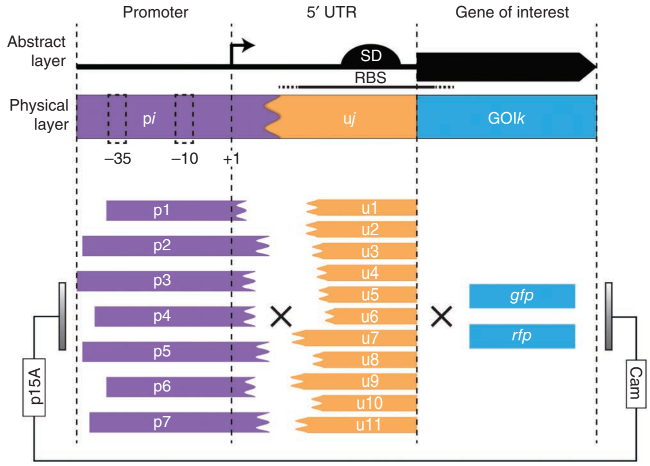


Figure 1 | Composition of irregular transcription and translation genetic elements. Schematic of 7 widely used promoters (p) and 11 5′ UTR (u) elements assembled in combination with two different genes of interest (GOIs), *gfp* and *rfp*, on a medium-copy (p15A) plasmid with chloramphenicol (Cam) resistance marker in *E. coli* (full element sequences via **Supplementary Table 1**). Promoters, 5′ UTRs and GOIs are typically considered to be well-defined, functionally independent genetic elements (abstract layer). However, irregular part boundaries create combination-specific junctions (physical layer) as parts are reused in combination (bottom). RBS, ribosome-binding site; SD, Shine-Dalgarno region.

We constructed a full combinatorial library of 7 promoters and 11 5′ UTRs upstream of two distinct genes of interest (*sfgfp*, hereafter *gfp*, and *mrfp1*, hereafter *rfp*; 52% nucleotide identity overall and 56% over the first 30 codons) and a common 3′ UTR context (Online Methods, **Fig. 1** and **Supplementary Fig. 1**). We placed each expression construct in a medium-copy-number plasmid and also integrated a subset of element combinations into the *E. coli* chromosome (Online Methods). We monitored expression of the different constructs via repeated measurements of steady-state fluorescence for both monomeric RFP (mRFP1, hereafter RFP) and superfolder GFP (sfGFP, hereafter GFP) and via analysis of individual transcripts using quantitative PCR (qPCR; Online Methods). Comparison of each measurement type enabled estimation of the individual contributions of transcription and translation processes to gene expression.

Measurement of promoter:5′ UTR combinations

We quantified gene expression across the combinatorial library by measuring fluorescence and mRNA levels under defined growth conditions (**Fig. 2a–d** and Online Methods). We monitored bulk culture fluorescence and growth profiles over time using an automated fluorometer; and we measured mRNA levels by qPCR and single-cell fluorescence distributions by flow cytometry, at a single time point during exponential growth (Online Methods). Single-cell and growth-normalized bulk-culture measurements of fluorescence were highly correlated ($R^2 = 0.96$ for both GFP and RFP libraries; **Supplementary Fig. 2a,b**) and exhibited high day-to-day reproducibility in triplicate experiments ($R^2 = 0.98$; **Supplementary Fig. 2c,d**). Fluorescence measurements were also well correlated between plasmids and chromosomal integrants ($R^2 = 0.85$; **Supplementary Fig. 3**). Once we established that no element combination produced bimodal expression distributions, we used average fluorescence levels as determined by cytometry for our analyses.

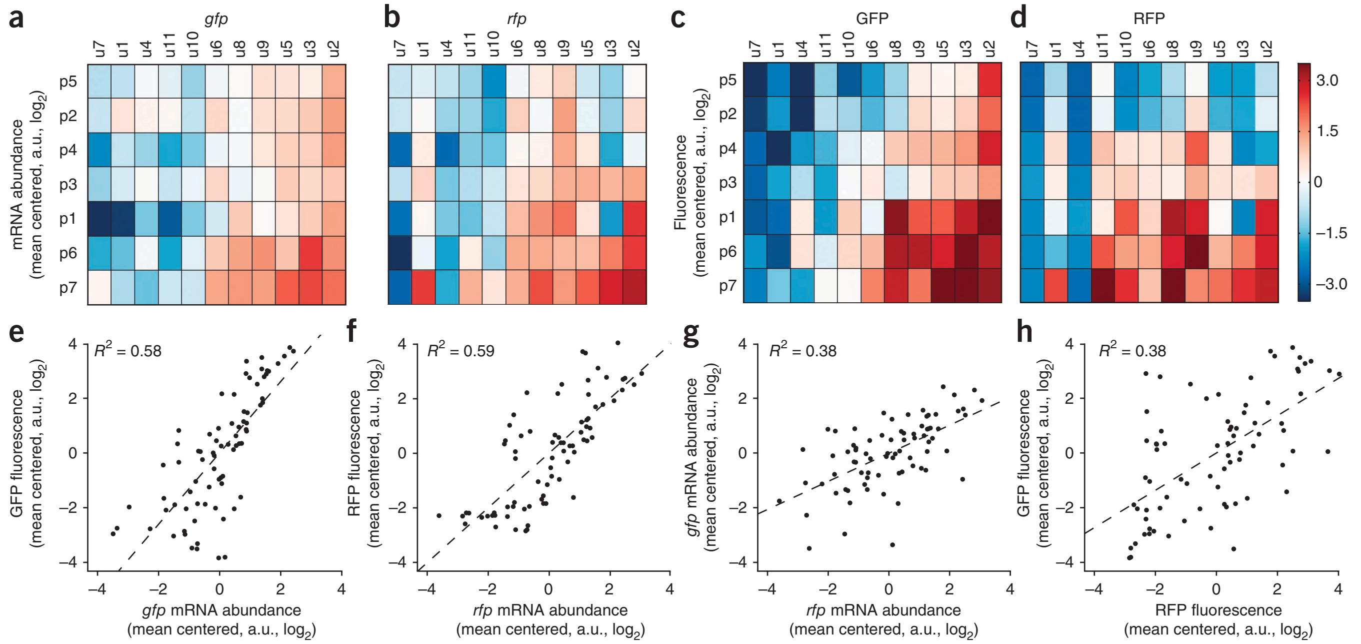


Figure 2 | Observed variation and correlation of mRNA abundance and protein fluorescence from a full combinatorial library of expression control elements. (**a**,**b**) Heat maps showing mRNA abundance for all combinations of transcription (p, rows) and translation (u, columns) elements driving the expression of *gfp* (**a**) or *rfp* (**b**). Each value is a dimensionless number corresponding to mean mRNA abundance measured from a cell population by bulk qPCR divided by the average abundance for all constructs within that panel. (**c**,**d**) Similarly mean-centered values for population average fluorescence intensities as measured by flow cytometry. The order of the elements in the matrices corresponds to a two-dimensional clustering performed on the data in **c** and held constant to facilitate visual comparison. Abundances are expressed on a $\log_2$ scale (mean-centered arbitrary units (a.u.)) and colored (thermometer scale). (**e**,**f**) mRNA abundance versus fluorescence for constructs driving *gfp* (**e**) and *rfp* (**f**) expression. (**g**,**h**) Pairwise comparison between mRNA levels (**g**) and fluorescence (**h**) for constructs driving *gfp* and *rfp* expression.

Variation and correlation of observed expression levels

Fluorescence values measured across the library varied over a 206- and 117-fold range for GFP and RFP, whereas mRNA levels varied over a 542- and 354-fold range, respectively. Protein and transcript abundance data indicated that a few promoters and 5′ UTRs encoded a consistent impact on expression across multiple part combinations (**Fig. 2**). Additionally, we observed some nonsystematic variation with specific combinations of promoters and 5′ UTRs across the two different reporters, indicating more complex interactions among parts. For example, the combination of promoter 1 and 5′ UTR 1 (p1:u1) produced ~11-fold more *rfp* than *gfp* mRNA and ~6-fold more red than green fluorescence (**Fig. 2g**,**h**). In contrast, the p1:u3 combination produced ~7-fold more *gfp* than *rfp* mRNA and ~37-fold more green than red fluorescence (**Fig. 2g**,**h** and **Supplementary Fig. 4**). We could not readily explain such differences by inspection of promoter core motifs (−35 spacer, −10 region), Shine-Dalgarno regions or total length of 5′ UTRs (**Fig. 2** and **Supplementary Table 1**).

A pairwise comparison of transcript abundance and fluorescence levels for each combination of control elements across the two reporters indicated that measured mRNA values accounted for ~60% of the total variation in fluorescence levels ($R^2 = 0.58$, $P < 2.2 \times 10^{-16}$ and $R^2 = 0.59$, $P < 2.2 \times 10^{-16}$ on a log-log scale for GFP and RFP libraries, respectively; **Fig. 2e**,**f** and **Supplementary Fig. 4a**,**b**). Pairwise comparison of transcript abundances ($R^2 = 0.38$, $P = 3.1 \times 10^{-9}$; **Fig. 2g** and **Supplementary Fig. 4c**) and fluorescence levels ($R^2 = 0.38$, $P = 3.1 \times 10^{-9}$; **Fig. 2h** and **Supplementary Fig. 4d**) between the two reporter libraries revealed more modest correlations.

Quantifying the performance of parts

We used a linear log-transformed model of gene expression (equation (3) in Online Methods) to quantify the individual contributions of promoters, 5′ UTRs and GOIs to different expression phenotypes (fluorescence, mRNA level or translation efficiency) and to quantify interactions among elements. More specifically, we conducted a full factorial ANOVA with repetitions[50] to predict the output of the system according to the identity of each element and element-element interactions as instantiated in any given construct.

Quantifying expected generic part type contributions

We first quantified how each category of genetic element and interactions among elements contribute to differences in expression. We assumed that the specific promoters, 5′ UTRs and GOIs used here define representative samples for each element type, and we used a random-effect interpretation of the ANOVA results[50]. From this assumption, we estimated the overall contribution of each element type and type-type interaction to expression levels (**Fig. 3** and **Supplementary Table 2**). We used mean-centered transcript abundances and fluorescence levels to remove confounding effects arising from systematic biases in experimental signals between GFP and RFP reporters while preserving interaction factors among the control elements and coding sequences themselves. We found that the 5′ UTRs and promoters are the major contributors to variation in expressed fluorescence levels (46% and 37%, respectively; **Fig. 3a** and **Supplementary Table 2a**). Also as expected[36,40], but quantified here, interactions across 5′ UTR: GOI junctions accounted for ~14% of total variation, whereas

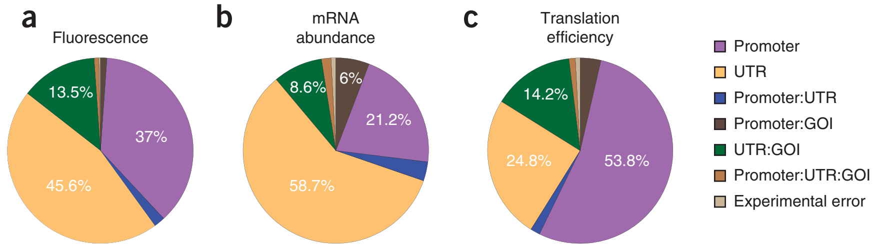


Figure 3 | Quantification of factors and interactions contributing to variation in mRNA abundance, translation efficiency and gene expression. Full factorial ANOVA[50] was conducted to quantify the average contributions from genetic element types, and from interactions among elements, with respect to total variation in measured gene expression levels. (**a**–**c**) Contributions of elements and interactions to total variation in protein fluorescence (**a**), mRNA abundance (**b**) and translation efficiency (**c**). 'Experimental error' represents the final term, ε, from equation (3) in Online Methods.

the combined contributions of all remaining interaction effects were negligible (<4% combined). Subsequent analysis found 5′ UTR identity to be the dominant factor (59%) in determining mRNA abundance, followed by the promoter (21%). Again, 5′ UTR:GOI interactions demonstrated an important contribution (9%) to mRNA abundance (**Fig. 3b** and **Supplementary Table 2b**). For translation efficiency, to our surprise, promoter identity emerged as the key factor (54%), followed by the 5′ UTR (25%) and 5′ UTR:GOI interactions (14%) (**Fig. 3c** and **Supplementary Table 2c**). The remaining error (ε) was the least important factor for all three experimental data types (fluorescence, mRNA and translation efficiency) and was well matched to experimental error, suggesting that unknown environmental factors or out-of-range measurements were not of concern and that a simple linear model is sufficient to explain observed variation in expression for the elements tested here.

Quantifying performance and quality for individual parts

We next quantified the primary activity of individual elements. We used a fixed-effect interpretation of the ANOVA results[50] to estimate part-associated scores that characterize the overall performance for any given genetic element and a set of subscores that quantify the variability in performance arising from interactions among elements (**Fig. 4**). Using the fluorescence data, we first estimated the main effect of each promoter, 5′ UTR and GOI to capture the average contribution of a given element to expression levels across all genetic contexts in which it was a component (Online Methods). We used these statistics to define a primary score for each element (**Fig. 4**). Primary scores must be corrected by an appropriate interaction term(s) to yield adjusted estimates of expression for a given combination of elements (equation (3) in Online Methods).

We then quantified variation in individual element activities across changing contexts as realized here by making many element-element combinations. We grouped element-element interactions by functional category and computed a set of secondary scores that define the sensitivity of a given element to each context variable (**Fig. 4** and Online Methods). The smaller a secondary score, the more likely it is that an element will maintain its primary activity across different contexts; larger secondary scores indicate greater context dependency. We found that, in general, the secondary scores relating 5′ UTRs to promoters were much smaller than those relating 5′ UTRs to GOIs. Thus, the secondary scores associated with promoters were typically small, indicating low context sensitivity for this class of functional elements as examined here (**Fig. 4**). Of note, elements with similar individual primary scores can have different secondary scores. For example, the u6, u11 and u10 5′ UTRs all had a similar average primary activity (~3) but different corresponding secondary scores with respect to changing GOI coding sequences (u11 > u10 > u6). In particular, the activity of u6 was largely insensitive to changes in either adjacent genetic element. Such information enables selection of elements for use in optimal testing strategies and in designing synthetic genetic systems.

We performed additional primary and secondary score analyses using mRNA abundance and translation efficiency data (**Supplementary Fig. 5a**,**b**). As with the fluorescence data, 5′ UTR:GOI junctions were found to be the largest source of variation in expression. The ranking of activity scores for each part was generally maintained across each data type (**Fig. 4** and **Supplementary Fig. 5**).

Predicting part performance

We determined whether and how our framework might be used to best estimate the expected average performance of new parts without having to construct and test all new possible part combinations. If there were no interactions among elements, then measurement of any new part within just a single combination of elements would be sufficient to perfectly estimate its performance across all elements. Alternatively, if each combination of elements produced highly specific effects, then all combinations might need to be assembled and tested.

Using the genetic elements studied here as a test set, we first observed how the quality of prediction for the expected average activity of a promoter increased as the number of 5′ UTRs with which it was tested increased (Online Methods). For example,

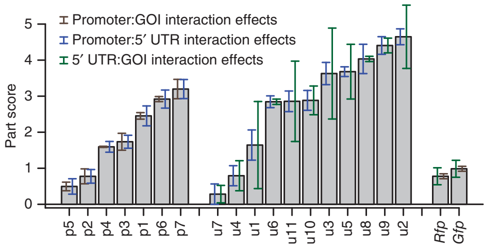
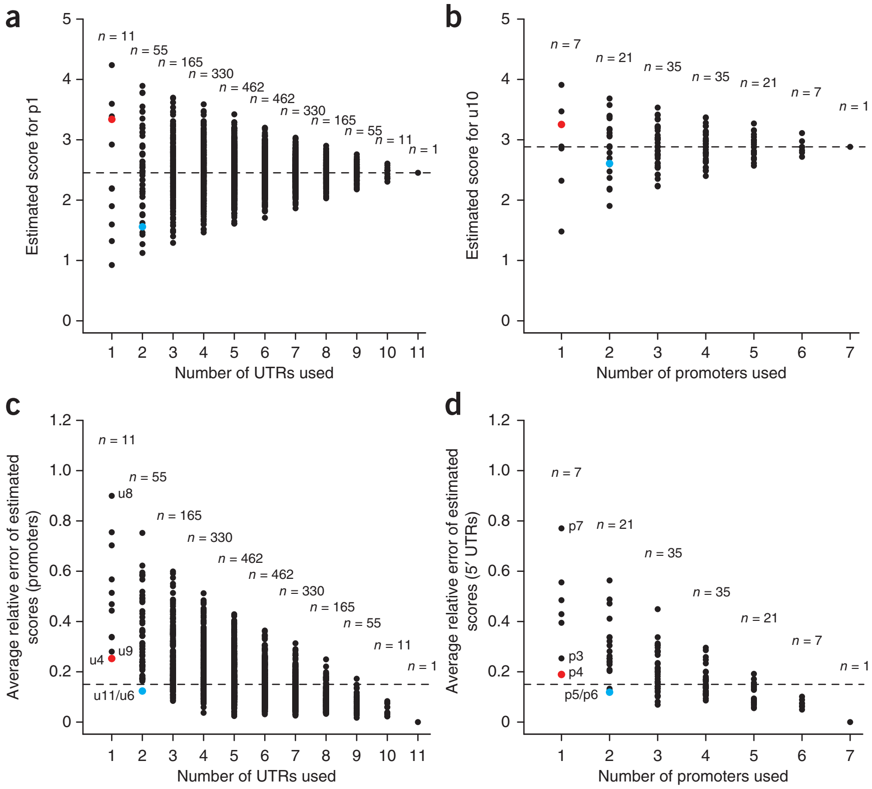


Figure 4 | Performance and quality scores for transcriptional and translation control elements. Primary part-activity scores (bar heights, log$_2$) giving the relative contribution of each promoter (p), 5′ UTR (u), and gene of interest (GOI) to observed fluorescence. Error bars indicate the standard error of all interactions involving each element with all other elements in a different functional category (Online Methods). As such, error bars reflect the variation of element performance in response to changes in proximal genetic context. Reciprocal interactions are color-coded as follows: gray, transcription elements and GOIs; blue, transcription and translation elements; green, translation elements and GOIs.

Figure 5 | Estimation of part activity with limited measurements. (**a**) Estimated activity for the promoter p1 with increasing numbers of 5′ UTRs. *n*, number of possible unique 5′ UTR combinations as a function of the number of 5′ UTRs tested. (**b**) Estimated activities of the 5′ UTR u10 with increasing numbers of promoters. (**c**) Relative error, averaged across all promoters, in estimating the activities of promoters with increasing numbers of 5′ UTRs (Online Methods). (**d**) Relative error, average across all 5′ UTRs, in estimating the activities of 5′ UTRs with increasing numbers of promoters. The individual parts (red) and part pairs (blue) that give the highest accuracy in estimating the activity of any new element are indicated.



estimating the quality of the most context-sensitive promoter, p1, using any single p1:5′ UTR combination produced an about eightfold range in estimated p1 activity (**Fig. 5a**). As the number of 5′ UTR combinations used to estimate p1 activity increased, the accuracy of the expected average p1 activity estimate improved. We observed similar trends for other promoters (**Supplementary Fig. 6a**) and also when we used one or more promoters to estimate the activity of 5′ UTRs (**Fig. 5b** and **Supplementary Fig. 6b**). Taken alone, these results might suggest that any new part must be tested with all other possible combinations of parts to estimate its expected average activity. Such work would quickly become prohibitively expensive as the number of parts and resulting part-part combinations increased. However, we noted that a few parts produced relatively accurate estimates of ensemble-average part activities even when just a single part combination was tested (**Fig. 5a**,**b**).

To better understand these observations, we computed the aggregate error in estimating the activities of all promoters (**Fig. 5c**) and 5′ UTRs (**Fig. 5d**) using varying numbers and combinations of 5′ UTRs and promoters, respectively (Online Methods). We found empirically that a limited number of measurements could be used to systematically estimate ensemble-wide part activities of promoters and 5′ UTRs with reasonable reliability. For example, the activity of any promoter could be estimated to within 15% of its average activity across all 5′ UTRs by using just two 5′ UTR measurements (u11 and u6). Similarly, the activity of any 5′ UTR could be determined to within 15% accuracy using just two promoter measurements (p5 and p6). Thus, for at least some biological element types, once a full combinatorial mapping is established for a particular interaction or context variable, much fewer experimental tests may be sufficient to accurately estimate the expected ensemble-average activity of new parts. Stated differently, after an initial seeding via brute force combinatorial measurements, increasingly more efficient, affordable and accurate part characterization can be realized via centralized facilities or distributed efforts.

Framework extension to an extrinsic environmental factor

We repeated GFP-only expression measurements in cultures grown at 30 °C to test whether the simple analysis framework developed here could account for an additional context variable. We found that GFP expression levels from the full combinatorial library were well correlated between 30 °C and 37 °C culture conditions ($R^2 = 0.97$; **Supplementary Fig. 7a**). Overall, changes across this temperature range accounted for less than 0.5% of total observed variation in GFP expressed from the given elements (**Supplementary Fig. 7b**), suggesting that the selected elements do not encode element- or junction-specific structures responsive to this temperature difference *per se*. We also noted that the overall promoter:5′ UTR interaction estimated from observed variation in 30 °C and 37 °C GFP levels was the same as that estimated from observed variation in 37 °C–only GFP and RFP levels (~2%, **Fig. 3a** and **Supplementary Fig. 7c**).

DISCUSSION

Our results demonstrate that the subtle functional couplings that arise as genetic elements are reused in combination can be systematically quantified and prioritized. Several advantages accrue from a systematic analysis spanning multiple element and junction types. For example, in recent work to normalize the 5′ UTR termini of mRNA by using ribozymes or a CRISPR-associated protein to cleave upstream sequences generated from irregular promoters, 'promoter-gene' couplings had been considered aggregate junctions[44,45]; splitting such a junction into all its functional elements (transcription, translation and GOI) might have instead allowed simple regularization of the +1 sequence encoded from the selected promoters[47] followed by separate work to normalize the junction spanning 5′ UTRs and downstream coding sequences. Stated differently, careful attention could ensure that irregular physical boundaries for elements selected from natural sources do not needlessly create complicating couplings. From such improvements, many approaches could then be tested to overcome remaining functional couplings involving more complicated element-element junctions. For example, following earlier observations[40] and confirmed by our analysis here, we determined that mRNA secondary structures at the 5′ UTR:GOI junction are correlated with expression levels (Pearson correlation coefficient (r) ≈ 0.8; **Supplementary Fig. 8**) and used this observation to motivate development and validation of an architecture for translation initiation elements whose activities are insensitive to changes in the coding sequence of downstream genes[47].

Although we can empirically determine costs required to estimate the expected activity of new parts for the types and conditions studied here (**Supplementary Table 3**), we can only predict the expected costs and recommend ways of organizing research for some types of future work. For example, we would expect that extending the work developed here, along with that presented in an accompanying manuscript[47], to another organism in which gene expression is controlled via similar molecular mechanisms (16S rRNA + 5′ UTR mRNA translation initiation) to equivalent precision (93% chance to realize factor of 2 change in expression ('factor-of-two levels') would require relatively little effort (**Supplementary Note**) and might be carried out in various laboratories simultaneously. Similarly, more parts for the types characterized here could likely be developed and tested in a distributed fashion by individual researchers. In contrast, we could not now quantify the effort required to increase expression reliability via parts that encode even more precise expression levels (for example: 99% chance to realize 'factor-of-2' levels); we would likely recommend that such work be conducted in a single facility. Similarly, systematic work with new types of parts, with new part junction architectures or under new physiological conditions should likely be closely coordinated such that innovations arising across individual laboratories can be rapidly complemented by medium-scale professional teams.

We caution that a simple linear model combined with ANOVA will likely not fully represent the activity and quality of genetic elements as increasing numbers of element types and environmental factors are considered simultaneously. The method also does not now track cell-cell variation for activities, resource utilization and many other important factors. Nevertheless, we have established that at least four key variables and their associated interactions can be observed, with subtle quantitative interactions affecting genetic element activity quantified given careful control of extrinsic physical variables. Extending even a simplistic approach to additional variables, genetic or environmental, will help to prioritize future work to improve parts[47]. Also, for environmental and physical factors found to affect part performance, we expect that additional work with *in vivo* reference materials will allow for a renormalization of part activities into functional spaces that are simpler to model[48]. Notably, such work can be carried out in an increasingly distributed and asynchronous fashion.

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 N. Hillson, V. Rhodius and C. Smolke for comments and discussion, and F. St-Pierre (Stanford University) for the plasmid pIT-KL-I52002. We acknowledge support from the National Science Foundation to the BIOFAB (award EEC 0946510). G.C. acknowledges support from the Human Frontier Science Program (LT000873/2011-l) and the Bettencourt Schueller foundation. A.P.A. and D.E. acknowledge support from the Synthetic Biology Engineering Research Center under National Science Foundation grant 04-570/0540879. J.C.G. acknowledges financial support from the Portuguese Fundação para a Ciência e a Tecnologia (SFRH/BD/47819/2008). This work was conducted at the Joint BioEnergy Institute supported by the Office of Science, Office of Biological and Environmental Research of the US Department of Energy, under contract DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

V.K.M., D.E. and A.P.A. conceived the study; V.K.M., G.C. and Q.-A.M. designed experiments; V.K.M., G.C., Q.-A.M., L.M., A.Y. and C.L. performed experiments; J.C.G. and G.C. built the computational model; V.K.M., G.C., J.C.G., D.E. and A.P.A. analyzed and interpreted the data; C.R. and M.J.C. provided software tools and database support; G.B. provided critical feedback on the framing the project; and V.K.M., J.C.G., G.C., J.D.K., D.E. and A.P.A. wrote the manuscript. All authors discussed and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Dubendorff, J.W. & Studier, F.W. Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with *lac* repressor. *J. Mol. Biol.* **219**, 45–59 (1991).
2. Mertens, N., Remaut, E. & Fiers, W. Tight transcriptional control mechanism ensures stable high-level expression from T7 promoter-based expression plasmids. *Bio/Technology* **13**, 175–179 (1995).
3. Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R. & Benenson, Y. Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science* **333**, 1307–1311 (2011).
4. Chen, Y.Y., Jensen, M.C. & Smolke, C.D. Genetic control of mammalian T-cell proliferation with synthetic RNA regulatory systems. *Proc. Natl. Acad. Sci. USA* **107**, 8531–8536 (2010).
5. Anderson, J.C., Clarke, E.J., Arkin, A.P. & Voigt, C.A. Environmentally controlled invasion of cancer cells by engineered bacteria. *J. Mol. Biol.* **355**, 619–627 (2006).
6. Saeidi, N. *et al.* Engineering microbes to sense and eradicate *Pseudomonas aeruginosa*, a human pathogen. *Mol. Syst. Biol.* **7**, 521 (2011).
7. Widmaier, D.M. *et al.* Engineering the *Salmonella* type III secretion system to export spider silk monomers. *Mol. Syst. Biol.* **5**, 309 (2009).
8. 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).
9. Ruder, W.C., Lu, T. & Collins, J.J. Synthetic biology moving into the clinic. *Science* **333**, 1248–1252 (2011).
10. Sinha, J., Reyes, S.J. & Gallivan, J.P. Reprogramming bacteria to seek and destroy an herbicide. *Nat. Chem. Biol.* **6**, 464–470 (2010).
11. Keasling, J.D. Manufacturing molecules through metabolic engineering. *Science* **330**, 1355–1358 (2010).
12. Carr, P.A. & Church, G.M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
13. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
14. Cambray, G., Mutalik, V.K. & Arkin, A.P. Toward rational design of bacterial genomes. *Curr. Opin. Microbiol.* **14**, 624–630 (2011).
15. Cardinale, S. & Arkin, A.P. Contextualizing context for synthetic biology—identifying causes of failure of synthetic biological systems. *Biotechnol. J.* **7**, 856–866 (2012).
16. Wilkinson, B. & Micklefield, J. Mining and engineering natural-product biosynthetic pathways. *Nat. Chem. Biol.* **3**, 379–386 (2007).
17. Canton, B., Labno, A. & Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.* **26**, 787–793 (2008).
18. Smolke, C.D. Building outside of the box: iGEM and the BioBricks Foundation. *Nat. Biotechnol.* **27**, 1099–1102 (2009).
19. Gulvanessian, H. & Holicky, M. Eurocodes: using reliability analysis to combine action effects. *Proceedings of the ICE - Structures and Buildings* **158**, 243–252 (2005).

20. 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).
21. 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).
22. Shimada, T. *et al.* Classification and strength measurement of stationary-phase promoters by use of a newly developed promoter cloning vector. *J. Bacteriol.* **186**, 7112–7122 (2004).
23. Zaslaver, A. *et al.* A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods* **3**, 623–628 (2006).
24. Babiskin, A.H. & Smolke, C.D. Synthetic RNA modules for fine-tuning gene expression levels in yeast by modulating RNase III activity. *Nucleic Acids Res.* **39**, 8651–8664 (2011).
25. Yarchuk, O., Jacques, N., Guillerez, J. & Dreyfus, M. Interdependence of translation, transcription and mRNA degradation in the *lacZ* gene. *J. Mol. Biol.* **226**, 581–596 (1992).
26. Cho, K.O. & Yanofsky, C. Sequence changes preceding a Shine-Dalgarno region influence *trpE* mRNA translation and decay. *J. Mol. Biol.* **204**, 51–60 (1988).
27. Telesnitsky, A.P.W. & Chamberlin, M.J. Sequences linked to prokaryotic promoters can affect the efficiency of downstream termination sites. *J. Mol. Biol.* **205**, 315–330 (1989).
28. Ellinger, T., Behnke, D., Knaus, R., Bujard, H. & Gralla, J.D. Context-dependent effects of upstream A-tracts - stimulation or inhibition of *Escherichia coli* promoter function. *J. Mol. Biol.* **239**, 466–475 (1994).
29. Stueber, D. & Bujard, H. Transcription from efficient promoters can interfere with plasmid replication and diminish expression of plasmid specified genes. *EMBO J.* **1**, 1399–1404 (1982).
30. Barrick, D. *et al.* Quantitative analysis of ribosome binding sites in *E.coli*. *Nucleic Acids Res.* **22**, 1287–1295 (1994).
31. Cox, R.S. III, Surette, M.G. & Elowitz, M.B. Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* **3**, 145 (2007).
32. Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. USA* **102**, 12678–12683 (2005).
33. Ellis, T., Wang, X. & Collins, J.J. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.* **27**, 465–471 (2009).
34. Reynolds, R. & Chamberlin, M.J. Parameters affecting transcription termination by *Escherichia coli* RNA: II. Construction and analysis of hybrid terminators. *J. Mol. Biol.* **224**, 53–63 (1992).
35. Carrier, T.A. & Keasling, J.D. Library of synthetic 5′ secondary structures to manipulate mRNA stability in *Escherichia coli*. *Biotechnol. Prog.* **15**, 58–64 (1999).
36. 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).
37. 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 (2012).
38. Khalil, A.S. *et al.* A synthetic biology framework for programming eukaryotic transcription functions. *Cell* **150**, 647–658 (2012).
39. Purnick, P.E. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.* **10**, 410–422 (2009).
40. de Smit, M.H. & van Duin, J. Control of translation by mRNA secondary structure in *Escherichia coli*. A quantitative analysis of literature data. *J. Mol. Biol.* **244**, 144–150 (1994).
41. Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C. & Wold, S. Quantitative sequence-activity models (QSAM)—tools for sequence design. *Nucleic Acids Res.* **21**, 733–739 (1993).
42. Yager, T.D. & von Hippel, P.H. A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*. *Biochemistry* **30**, 1097–1118 (1991).
43. 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).
44. 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).
45. 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).
46. Klumpp, S., Zhang, Z. & Hwa, T. Growth rate-dependent global effects on gene expression in bacteria. *Cell* **139**, 1366–1375 (2009).
47. Mutalik, V.K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* advance online publication, doi:10.1038/nmeth.2404 (10 March 2013).
48. Kelly, J.R. *et al.* Measuring the activity of BioBrick promoters using an *in vivo* reference standard. *J. Biol. Eng.* **3**, 4 (2009).
49. Kittleson, J.T., Wu, G.C. & Anderson, J.C. Successes and failures in modular genetic engineering. *Curr. Opin. Chem. Biol.* **16**, 329–336 (2012).
50. Wu, C.F.J. & Hamada, M.S. *Experiments: Planning, Analysis, and Optimization*, 2nd edn (Wiley, Hoboken, New Jersey, USA, 2009).

ONLINE METHODS

Bacterial strains, plasmids and growth conditions. Plasmids and strains used in this study are listed in **Supplementary Table 4**. Oligonucleotides used in this work are listed in **Supplementary Table 5**. Detailed information on sequence-based design, plasmid maps and corresponding experimental data sets for each construct are available in a public repository database at http://biofab.org/data.

All plasmid manipulations were performed using standard cloning techniques[51]. All enzymes used for plasmid manipulations were obtained from New England Biolabs (NEB), and oligonucleotides were ordered from Integrated DNA Technologies (IDT). *E. coli* strain BW25113 was used for plasmid construction purposes and for fluorescence measurements. All strains were grown in MOPS EZ Rich Medium (Teknova) supplemented with 34 μg/ml chloramphenicol at 37 °C (or 30 °C), with shaking at 900 r.p.m. All of the experiments were conducted in triplicate (replicate assays from independent overnight liquid cultures).

Promoter:5′ UTR combinatorial library construction. For assembling the promoter:5′ UTR combinatorial library, we modified the Golden Gate method[52] to comply with assembly of smaller (<80 bp) promoter and 5′ UTR fragments as annealed oligonucleotides and not as plasmid-borne fragments as typically recommended[52]. The compatibility of overhangs of fragments to be assembled was confirmed using j5 software[53].

The promoter:5′ UTR combinatorial library driving the expression of GFP[54] and RFP[55] from the reporter genes *gfp* and *rfp* were carried on the medium-copy vectors pFABOUT2 and pFABOUT18, respectively. pFABOUT2 and pFABOUT18 were derived from the same backbone vector pBbA2c-RFP[56] and have a TetR-regulated Ptet promoter driving the expression of either reporter, a p15A replication origin and chloramphenicol resistance marker (**Fig. 1** and **Supplementary Fig. 1**).

To prepare the backbone vector for assembly of the promoter and 5′ UTR library, we amplified pFABOUT2 using primers oFAB57 and oFAB58 (**Supplementary Table 5**) with Phusion high-fidelity DNA polymerase (NEB, manufacturer's instructions). The forward primer introduces a BsaI recognition site such that after restriction digestion of the PCR products, we obtained an overhang of 5′-ATGA-3′ (**Supplementary Fig. 1a**). Similarly, the reverse primer introduces a BsaI site such that we obtained an overhang 5′-GATA-3′ on the restriction-digested vector backbone. The PCR amplification also eliminated the *tetR* gene, Ptet promoter and 5′ UTR driving *gfp* from pFABOUT2 (**Supplementary Fig. 1**). The PCR-amplified backbone products were purified using PCR purification kit (Qiagen), digested with BsaI enzyme (37 °C, overnight (>18 h)) and purified again using a PCR purification kit (Qiagen) to yield pFABOUT2_cut. Similarly, pFABOUT18, bearing the reporter gene *rfp*, was amplified with primers oFAB58 and oFAB60 to introduce BsaI recognition sites, and purified PCR products were digested with BsaI to yield pFABOUT18_cut (**Supplementary Fig. 1b**).

The 7 promoters and 11 5′ UTRs (and 1 Null-RBS 5′ UTR as a control) were chosen from various literature sources (**Supplementary Table 1**), and these elements were each assembled by annealing corresponding forward and reverse oligonucleotides (**Supplementary Table 5**). The forward and reverse oligonucleotides for 5′ UTR and promoter regions were designed to yield overhangs compatible with each other and with the digested vector backbones (**Supplementary Fig. 1**) and were confirmed using the software j5 (ref. 53). To minimize the sequence constraints on the 5′ UTR, we chose the start codon region of the reporter as an overhang that is compatible with the 5′ UTR overhangs (**Supplementary Fig. 1**). Because the fourth nucleotide in the *gfp* and *rfp* genes is different, we designed different reverse primers for 5′ UTRs depending on the destination vector backbones. The phosphorylated overhang TCAT at the 3′ end of annealed 5′ UTR part is compatible with the ATGA at the 5′ end of the *gfp* gene within a pFABOUT2_cut vector backbone (**Supplementary Fig. 1a**). Similarly, the phosphorylated overhang CCAT at the 3′ end of an annealed 5′ UTR part is compatible with the ATGG at the 5′ end of the *rfp* gene within a pFABOUT18_cut vector backbone (**Supplementary Fig. 1b**). The phosphorylated TTTG overhang at the 5′ end of an annealed 5′ UTR part is compatible with the CAAA at the 3′ end of annealed promoter parts, whereas the phosphorylated TATC overhang at the 5′ end of annealed promoter parts is compatible with the 5′ GATA overhang of the restriction-digested vector backbone.

The forward and reverse oligonucleotides (received from IDT, **Supplementary Table 5**) were reconstituted in sterile water to make up 100 μM stocks. The phosphorylation of oligonucleotides was performed using T4 polynucleotide kinase (PNK from NEB) in T4 ligase buffer (with ATP). Annealing reactions were performed by incubating the phosphorylated complementary oligonucleotides at 95 °C for 3 min (5 μl of 100 μM forward and reverse oligonucleotides in 90 μl of sterile water) and followed by cooling at room temperature for 30 min. The phosphorylated-annealed oligonucleotides were then ligated to the digested PCR products, pFABOUT2_cut and pFABOUT18_cut using T4 DNA ligase in 96-well PCR plates and transformed into chemically competent BW25113 cells (in-house prepared in 96-well plates). Positive clones were then confirmed by PCR amplification and sequencing of the assembly region by using primers soFAB8 and soFAB23. The overnight cultures of positive clones were stored in glycerol stocks as per standard procedure[51]. The combinatorial assembly of promoters and 5′ UTRs thus generated a total 168 plasmids comprising 84 constructs of the GFP library and 84 constructs of the RFP library.

Insertion of the combinatorial library on the chromosome. A subset of the library driving the expression of *gfp* comprising each of the 7 promoters combined to at least 6 of the 11 5′ UTRs was inserted on the chromosome at the main target attachment site of the phage lambda (*attBλ*). The expression cassette comprising 43 bp upstream of the promoter to the end of the terminator was amplified by PCR using primers oFAB174 and oFAB175 from the plasmids described above, resulting in 986-bp amplicons. The PCR products were purified, digested with XbaI and PstI, purified again and ligated into pFABOUT16 digested by XbaI and NsiI. pFABOUT16 is a derivative of pIT-KL-I52002 (gift from F. St-Pierre; with D.E., unpublished data). This plasmid contains the conditional origin of replication R6Kγ, which requires the product of the *pir* gene to be functional. Thus, the ligation products were transformed into chemocompetent *E. coli* EC100D (Epicentre) and selected for kanamycin resistance. The validity of each clone was confirmed by sequencing, using oligonucleotides

doi:10.1038/nmeth.2403

soFAB15 and soFAB16. pFABOUT16 contains the sequence of the lambda phage attachment site (*attPλ*). Validated clones were transformed into the induced chemocompetent BW25113 strain carrying the helper plasmid pLambdaInt[57]. The plasmid pLambdaInt expresses the lambda recombinase under the control of a thermosensitive promoter induced at 37 °C. Because the plasmid must integrate in a replicon to be propagated in a *pir-* strain, *attPλxattBλ* recombinants were selected by screening for kanamycin resistance. The correct positioning and copy number of the integrated plasmid were confirmed by PCR using primers soFAB17/soFAB19 and soFAB17/soFAB18, respectively, as described[56]. Fluorescence levels were measured by flow cytometry using the same standard protocol as described below.

***In vivo* assays using the plate reader and flow cytometer.** Assay strains were stored as glycerol stocks in 96–deep-well plates (2 ml) and in smaller aliquots of 50 µl in 96-well sterile PCR plates for easy handling. These aliquots were used only for triplicate experiments and were then discarded to avoid any unwanted physiological changes due to repeated freeze-thaw cycles. We grew cultures in 2–ml–, 96–deep-well plates containing 400 µl of MOPS EZ Rich Medium (Teknova, cat. #M2105) with appropriate antibiotics, inoculating 3 µl from thawed glycerol stocks. Cultures were grown overnight (~16 h) in plates of 96 U-shaped 2-ml wells covered with sterile breathable sealing film at 37 °C with shaking at 900 r.p.m. on a Multitron shaker (Infors-HT). We note here that promoter p5 (pLlacO1) encodes a *lac* operator site[58], and because *E. coli* strain BW25113 is $lacI^+$ (ref. 59), promoter p5 is partially repressed under our experimental conditions.

For microplate kinetic assays, overnight cultures were diluted 1:50 into a final volume of 150 µl of fresh medium with appropriate antibiotics in clear-bottom black plates and incubated in a multimode microplate reader-incubator-shaker Synergy-2 (BioTek Instruments). Cultures were grown for 6 h with rapid shaking and repeated measurements for optical density (OD_{600}) and fluorescence (relative fluorescence units or RFU; 481-nm excitation and 507-nm emission for GFP; 560-nm excitation and 650-nm emission for RFP) were performed every 10 min. All microplate kinetic assay experiments were repeated at least three times starting from independent overnight cultures. Gen5 software for BioTek plate readers was used for data acquisition, and further data analysis was performed using MATLAB software (MathWorks) with in-house–developed scripts (see below).

For the flow cytometer assays, overnight cultures were diluted 1:50 into a final volume of 200 µl fresh 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 the Multitron shaker. Cultures were diluted 1:2,000 in chilled and filtered PBS (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 microcapillary and autosampling capabilities, 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. This protocol was set to minimize any extrinsic source of variation. Guavasoft software was used for data acquisition, and the resulting FCS files were further analyzed using in-house–developed R scripts (see below; available upon request).

Transcriptional analysis by qPCR. Cultures were prepared and grown as described above for 2 h to reach exponential phase with OD_{600} in the range of 0.3–0.5 when measured in a microplate reader. After 2 h, cultures were harvested on ice and total RNA was extracted by enzymatic lysis with lysozyme, which was followed by β-mercaptoethanol and ethanol treatment. RNA cleanup procedure was performed using the Qiagen 96 RNA Protect Kit and vacuum manifold (Qiagen); samples were eluted in 120 µl of RNase-free water and stored at −80 °C. Total RNA concentration in each sample of 96-well plate was quantified using a Nanodrop 1000 (Thermo Scientific), and a volume corresponding to 25 µg of total RNA was used for qPCR. Using the Power SYBR Green RNA-to-C_T 1-Step Kit (Applied Biosystems), we performed reverse transcription of RNA standard and samples, immediately followed by qPCR in a one-step reaction in Applied Biosystem's StepOnePlus instrument (manufacturer's protocol). The transcript abundances of reporter genes *gfp* and *rfp* were quantified using a standard curve.

To establish the standard curve for *gfp* and *rfp* transcripts, we PCR-amplified each gene using oFAB1360/oFAB1364 from pFABOUT2 and oFAB1459/oFAB1460 from pFABOUT18, respectively. Gel electrophoresis and Bioanalyzer 2100 (Agilent) assays were used to confirm the amplification, and purified products were directly used as template for *in vitro* transcription with the MEGAscript T7 Kit (Ambion) according to the manufacturer's protocol. Completed transcription reactions were followed by DNase I treatment (TURBO DNase, Ambion). After determination of RNA concentration by spectrophotometry (Nanodrop 1000, Thermo Scientific) and Bioanalyzer 2100 (Agilent), the copy numbers of standard RNA molecules were calculated using the following formula: (X g/µl RNA / (transcript length in nucleotides × 340)) × 6.022 × 10^{23} = Y molecules per µl. A dilution series ranging from 10^{11} to 10^4 molecules per µl of each template was prepared; aliquots were made and then stored at −80 °C. Each template (1 µl) was used for reverse transcription and qPCR according to the protocol described above.

Plate reader kinetic assay data analysis. Background fluorescence of cultures was determined using a combinatorial library of seven promoters combined with a nonfunctional 5′ UTR ('dead RBS (Null RBS)' 5′ UTR) with each of the two reporters in *E. coli* BW25113 (**Supplementary Table 1**). These control strains were grown and assayed along with each of the combinatorial libraries on the same 96-well plates. Their fluorescence signals were averaged to generate a standard curve for OD against fluorescence (RFU). The standard curve was used to subtract background fluorescence from the reporter strain fluorescence value at the same OD, yielding a background subtracted OD vs. RFU differential rate plot for each strain carrying each member of the combinatorial library[20]. The slope of the linear portion of each differential rate plot was taken as exponential steady-state fluorescence (f_{ss}). To account for the growth rate and the maturation rate

doi:10.1038/nmeth.2403

constant for the stable fluorescent protein, we used a previously published model[60]

$$\text{Expression strength} = f_{\text{ss}} \times \mu \times \left(1 + \frac{\mu}{m}\right) \qquad (1)$$

where μ is the culture growth rate, m is the fluorescent reporter maturation constant and f_{ss} is the steady state fluorescence. The maturation rates for GFP and RFP were obtained from the literature (~2.77 h^{-1} for GFP[61] and ~1 h^{-1} for RFP[55]), whereas both steady-state fluorescence (relative nonfluorescent units per OD) and growth rate (h^{-1}) were experimentally determined for each candidate in the library. The expression strength estimated by this procedure yields the synthesis rate of nonfluorescent protein (GFP or RFP) $\text{OD}^{-1}\ \text{h}^{-1}$ (ref. 60), which was used for comparison with the exponential-phase single-cell mean fluorescence measured by flow cytometry.

Flow cytometer data analysis. For each replicate, the FCS files were parsed and analyzed using in-house scripts for R software (http://www.r-project.org/). The gating parameters defined at the time of acquisition were reapplied. In addition, a custom automated gating procedure was developed to maximize consistency in the results[62]. The measurements were first filtered through an ellipsoid gate set around the main bicluster of log–forward- and log–side-scatter data. The resulting data were then clustered on the appropriate log-fluorescence signal, allowing for one or two clusters. This step was used to control the quality of the data through the identification of well-to-well contaminations or selection for loss–of-function mutants, both of which occasionally occurred during experiments. A specific criterion combining the integrated completed likelihood (ICL) and the Bayesian information criterion (BIC) for the fitted mixture model was applied to determine the presence of bimodality in the fluorescence data. We used this strategy to flag experiments that needed to be redone. In such cases, a dilution of the population was plated and individual colonies were then screened and resequenced. A validated isolate was used to rerun the experiment on both the plate reader and flow cytometer.

For each filtered data set, the average and variance of the appropriate linear fluorescence data was calculated, yielding a relative measure of fluorescence per cell (RFU per cell, **Supplementary Fig. 2a**,**b**). The calculated averages are highly correlated to the bulk plate reader data. The fluorescence-per-cell values were used for all the analysis developed in this work.

ANOVA models for fluorescence, mRNA abundance and translation efficiency. We first considered a conventional model[46] to represent population average steady-state protein expression levels from a constitutive promoter and ribosome-binding site (RBS)

$$P = g \times \text{Tx} \times \text{Tr} / (\text{kd} \times \text{km}) \qquad (2\text{a})$$

$$\log(P) = \log(g) + \log(\text{Tx}) + \log(\text{Tr}) - \log(\text{km}) - \log(\text{kd}) \qquad (2\text{b})$$

where P is a steady-state protein level, g is the gene copy number, Tx the transcription rate per gene copy, Tr the translation rate per mRNA, and km and kd are the mRNA and protein degradation rates, respectively. Typically, such models (equation (2)) assume that the individual activities of genetic elements are fixed within a given context. However, if elements are reused in novel combinations or across varying operational contexts, then their activities can change. We thus developed a linear model inspired from equation (2b) to enable analysis of element-element context effects.

To understand the contribution and coupling between transcription and translation elements (promoters and 5′ UTRs) and divergent GOIs on overall gene expression, we consider the following equations for ANOVA

$$\begin{aligned}\log(O_{ijk}) = \alpha + P_i + U_j + \text{GOI}_k + (P{:}U)_{ij} + (P{:}\text{GOI})_{ik} \\ + (U{:}\text{GOI})_{jk} + (P{:}U{:}\text{GOI})_{ijk} + \varepsilon_{ijk} \\ \text{for } i = \{1-7\}; j = \{1-11\}; k = \{1,2\} \qquad (3)\end{aligned}$$

where O_{ijk} is an expression output signal (for example, arbitrary fluorescence level, transcript abundance and translation efficiency or fluorescence per mRNA) measured from a genetic construct comprising a member of combinatorial promoter:5′ UTR with either GFP or RFP as a reporter; P_i represents the ith promoter; U_j represents the jth 5′ UTR; GOI_k represents the kth reporter; $(P{:}U)_{ij}$ represents the interaction between the ith promoter and jth 5′ UTR; $(P{:}\text{GOI})_{ik}$ represents the interaction between the ith promoter and kth reporter; $(U{:}\text{GOI})_{jk}$ represents the interaction between the jth 5′ UTR and kth reporter; $(P{:}U{:}\text{GOI})_{ijk}$ represents the interaction between the ith promoter, jth 5′ UTR and kth reporter; α is the overall average signal; and the term ε_{ijk} represents the error term for the ith promoter, jth 5′ UTR and kth reporter combination.

Relationship between model entities and biophysical mechanisms. There may exist a complex mapping between various factors considered in equation (2) and equation (3). We thus log-transformed our experimental data sets and represented the relationship as an additive form of a linear model. Similar approaches have been applied in analyzing gene expression via microarray data sets[63,64]. Stated differently, we assumed that log-normalized gene expression is a linear function of different factors and their interactions, and each factor is an abstraction of complex biophysical functions encoded in primary DNA sequences. For instance, a given factor P_i encompasses the contribution of promoter sequence motif recognition by RNA polymerase (RNAP) and also multiple steps involved in transcription initiation and subsequently promoter escape[21]. Although each of these parameters is known to affect the activity of a promoter, they are lumped here into a single estimated value, P_i. Concurrently, promoter elements may also contribute a few nucleotides to the actual 5′ UTR of resulting transcripts and can therefore affect all mechanisms depending on the mRNA molecule directly (for example, transcript stability and translation initiation). Furthermore, a promoter downstream region can comprise the 5′ end encoded from any given 5′ UTR, which may affect RNAP pause and promoter escape[21]. Such complex relationships are captured by the $P{:}U_{ij}$ interaction terms. Similarly, a given U_j captures contributions due to ribosome binding and mRNA stabilization, which affects rates of both translation initiation and transcript degradation, whereas its interaction term with GOI, $(U{:}\text{GOI})_{jk}$ describes the differential impact of different 5′ UTRs on translation initiation of any downstream gene (for example, interfering with translation

doi:10.1038/nmeth.2403

initiation by inhibitory structures specific to different GOIs or by modifying transcript stability[65]).

The factor GOI_k in equation (3) also captures intrinsic differences in translation elongation properties of GOIs (for example, translation pause, codon usage effects and protein maturation), protein degradation, and fluorescence intensity itself. This variable also captures measurement artifacts such as differential efficiency of qPCR reactions across various templates for mRNA quantification as well as differences in fluorescence intensities linked to the use of different optical filters sets between reporters. The interaction term $(P{:}\mathrm{GOI})_{ik}$ accounts for possible interactions between a promoter, P, and a downstream GOI (for example, a gene may have an internal promoter that interferes with the RNAP loading from an upstream promoter or cause transcriptional interference[66]). The higher-order interaction term $(P{:}U{:}\mathrm{GOI})_{ijk}$ does not relate to any known example or mechanism for how variation in gene expression is thought to arise from specific functional genetic elements (for example, promoters, 5′ UTRs and GOIs) and hence captures any factors not included in the model.

Replicates. The term ε_{ijk} in equation (3) accounts for experimental errors arising from the replicate data sets or systematic experimental biases (for example, signal saturation). In the ANOVA, this error term is assumed to be independent and to have zero mean. In the present work measurements were performed in batches of 96-well plates (three different plates for the GFP library and three different plates for the RFP library). ε_{ijk} was calculated by performing ANOVA using equation (3), and the results indicated insignificant impacts (see Results).

To formally check for systematic block effects and validate our treatment of the replicates, we performed a three-way ANOVA[50] without replicate data sets on each of the libraries according to the following linear model

$$\log(O_{ijk}) = \alpha + P_i + U_j + R_k + \varepsilon_{ijk}$$
$$\text{for } i=\{1-7\};\ j=\{1-11\};\ k=\{1-3\} \qquad (4)$$

where R_k stands for replicate batch number, and all other notations have the same meaning as in equation (3). We found no contribution of the replicate factor on fluorescence data sets (F-statistic scores = 0.431 and 0.388, P = 0.651 and 0.679 for GFP and RFP fluorescence data, respectively).

Sum of squares and score calculations. Equations (5–7) relate fluorescence (F), transcript abundance (M) and fluorescent protein produced per mRNA (TE, translation efficiency) to promoters, 5′ UTRs and GOIs used in our combinatorial library. In each regression model, the colon denotes an interaction between different elements, and 'Exp. Error' is the variation resulting from measurement replicates.

$$\begin{aligned}\log(F) = \alpha &+ \text{promoter} + \text{UTR} + \text{GOI} + \text{promoter:UTR}\\ &+ \text{promoter:GOI} + \text{UTR:GOI}\\ &+ \text{promoter:UTR:GOI} + \text{Exp. Error}\end{aligned} \qquad (5)$$

$$\begin{aligned}\log(M) = \alpha &+ \text{promoter} + \text{UTR} + \text{GOI} + \text{promoter:UTR}\\ &+ \text{promoter:GOI} + \text{UTR:GOI}\\ &+ \text{promoter:UTR:GOI} + \text{Exp. Error}\end{aligned} \qquad (6)$$

$$\begin{aligned}\log(\mathrm{TE}) = \alpha &+ \text{promoter} + \text{UTR} + \text{GOI} + \text{promoter:UTR}\\ &+ \text{promoter:GOI} + \text{UTR:GOI}\\ &+ \text{promoter:UTR:GOI} + \text{Exp. Error}\end{aligned} \qquad (7)$$

To account for differences in the fluorescence intensities of reporters and in qPCR primer efficiency, we mean-centered our data sets for each GOI independently. Using three replicates of fluorescence, transcript abundance and translation efficiency, we performed ANOVA[50] on the linear models (equations (5–7)) described above using the "anova" routine in R software. ANOVA results are presented in **Figures 3** and **4** and in **Supplementary Table 2**.

The main effects (the primary scores for promoters, 5′ UTRs and GOIs) were directly retrieved from the ANOVA table of effects (accessed using the "model.tables" function in R). In essence, scores are calculated as the mean of all observations comprising a given level of a factor from which the grand mean (α) is subtracted (for example, the mean of all observations containing p1 promoter subtracted by the mean of all observations to yield the primary score for p1). For ease of visualization and interpretation, the grand mean (α) was distributed over the main effects so that all resulting scores were positive.

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 (for example, the context sensitivity of p1 due to composition with different UTRs was calculated using the s.e.m. of the set {p1:u1 effect, p1:u2 effect, …, p1:u11 effect}). By definition, the sum of these interaction terms equals 0. Hence, the s.e.m. effectively measures the spread of the interaction around main effects.

Accuracy of estimated scores for new parts. To evaluate the accuracy of our model for use in estimating the scores for a new promoter and/or 5′ UTR (i.e., for elements not included in the training model), we performed a cross-validation analysis using our entire data set. Here, the goal is to determine how the robustness of an estimated score for a new promoter (or 5′ UTR) changes as we characterize this new promoter (or 5′ UTR) with an increasing number of 5′ UTRs (or promoters). Please note that here we define the 'true' score of a promoter (or 5′ UTR) as its estimated score across all combinations (i.e., across the entire data set, as presented above).

In detail, the following steps were taken to determine the accuracy of the estimated score for a new promoter by a cross-validation approach:

1. Create different combinations of 5′ UTRs with a varied number of elements ranging from 1 to 11 (e.g., (u1), …, (u11), (u1, u2), …, (u10, u11), (u1, u2, u3), …).

2. For each of combination of 5′ UTRs defined in step (1), perform the following. (i) Estimate the score of each of the seven promoters in the data set using the experimental data corresponding to only the combination of these promoters with the 5′ UTR elements present in that combination and based on the model in equation (5). (ii) Calculate the absolute error (AE) as the difference between the estimated promoter score (step (i)) and the true score of the promoter (estimated with all combinations of 5′ UTRs). (iii) Calculate the relative error (RE) by dividing the AE by the true score of the promoter. (iv) Average all the

doi:10.1038/nmeth.2403

relative errors across all the promoters to produce an aggregated metric for that number of test 5′ UTR combinations. That is, for every 5′ UTR group defined in step (1), we calculated seven REs (because we have seven promoters), averaged them all to produce a single metric and used this score to generate the plot shown in **Figure 5**.

An equivalent analysis was carried out to determine the accuracy of estimating the score of a new 5′ UTR as a function of the number of combinations of promoters tested.

Evaluation of temperature effect on parts performance. To estimate the effect of temperature on the performance of transcription and translation elements (promoter and 5′ UTR), we consider the following equations for ANOVA

$$\begin{aligned}\log(F_{ijk}) &= \alpha + P_i + U_j + T_k + (P{:}U)_{ij} + (P{:}T)_{ik} \\ &\quad + (U{:}T)_{jk} + (P{:}U{:}T)_{ijk} + \varepsilon_{ijk} \\ &\quad \text{for } i = \{1-7\};\ j = \{1-11\};\ k = \{1, 2\}\end{aligned} \tag{8}$$

where F_{ijk} is the fluorescence level measured from a genetic construct comprising a member of combinatorial promoter:5′ UTR grown at either 30 °C or 37 °C; P_i represents the ith promoter; U_j represents the jth 5′ UTR; T_k represents the kth temperature; $(P{:}U)_{ij}$ represents the interaction between ith promoter and jth 5′ UTR; $(P{:}T)_{ik}$ represents the effect of the kth temperature on the ith promoter; $(U{:}T)_{jk}$ represents the effect of the kth temperature on the jth 5′ UTR; $(P{:}U{:}T)_{ijk}$ represents the interaction between the ith promoter, jth 5′ UTR and kth temperature; α is the overall average signal; and ε_{ijk} represents the error term for the ith promoter, jth 5′ UTR and kth temperature combination.

51. Ausubel, F.M. *Short Protocols in Molecular Biology*, 5th edn (Wiley, New York, 2002).
52. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).
53. Hillson, N.J., Rosengarten, R.D. & Keasling, J.D. j5 DNA assembly design automation software. *ACS Synth. Biol.* **1**, 14–21 (2012).
54. 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).
55. Campbell, R.E. *et al.* A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. USA* **99**, 7877–7882 (2002).
56. Lee, T.S. *et al.* BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *J. Biol. Eng.* **5**, 12 (2011).
57. Haldimann, A. & Wanner, B.L. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *J. Bacteriol.* **183**, 6384–6393 (2001).
58. 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).
59. 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).
60. Leveau, J.H. & Lindow, S.E. Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J. Bacteriol.* **183**, 6752–6762 (2001).
61. Iizuka, R., Yamagishi-Shirasaki, M. & Funatsu, T. Kinetic study of *de novo* chromophore maturation of fluorescent proteins. *Anal. Biochem.* **414**, 173–178 (2011).
62. Lo, K., Hahne, F., Brinkman, R.R. & Gottardo, R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* **10**, 145 (2009).
63. Kerr, M.K. & Churchill, G.A. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201 (2001).
64. Kerr, M.K., Martin, M. & Churchill, G.A. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**, 819–837 (2000).
65. Ringquist, S. *et al.* Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol. Microbiol.* **6**, 1219–1229 (1992).
66. Shearwin, K.E., Callen, B.P. & Egan, J.B. Transcriptional interference—a crash course. *Trends Genet.* **21**, 339–345 (2005).

doi:10.1038/nmeth.2403