GENETICS

Deep-learning–assisted Sort-Seq enables high-throughput profiling of gene expression characteristics with high precision

Huibao Feng[1]†*, Fan Li[1]†, Tianmin Wang[2,3], Xin-hui Xing[1,4], An-ping Zeng[5,6], Chong Zhang[1,7]*

Owing to the nondeterministic and nonlinear nature of gene expression, the steady-state intracellular protein abundance of a clonal population forms a distribution. The characteristics of this distribution, including expression strength and noise, are closely related to cellular behavior. However, quantitative description of these characteristics has so far relied on arrayed methods, which are time-consuming and labor-intensive. To address this issue, we propose a deep-learning–assisted Sort-Seq approach (dSort-Seq) in this work, enabling high-throughput profiling of expression properties with high precision. We demonstrated the validity of dSort-Seq for large-scale assaying of the dose-response relationships of biosensors. In addition, we comprehensively investigated the contribution of transcription and translation to noise production in *Escherichia coli*, from which we found that the expression noise is strongly coupled with the mean expression level. We also found that the transcriptional interference caused by overlapping RpoD-binding sites contributes to noise production, which suggested the existence of a simple and feasible noise control strategy in *E. coli*.

INTRODUCTION

Gene expression is essential in transmitting genetic information from genes to RNA and then to proteins within each cell, influencing its characteristics. However, gene expression is often stochastic, as it involves many random events requiring the participation of various low–copy-number chemical components (*1*–*6*). In addition, this process can be chaotic because of the high complexity of the regulatory network (*7*, *8*). As a result, phenotypic heterogeneity exists among genetically identical cells even under the same environmental conditions (*1*). Therefore, steady-state protein production in a clonal population exhibits a distribution, wherein the mean of the distribution (mean) indicates the expression strength, and the squared coefficient of variation (CV^2) exhibits the expression noise. These two characteristics are both important indicators that are closely related to the phenotypes of a population, such as bioproduction efficiency (*9*, *10*), drug resistance (*11*, *12*), and antibiotic persistence (*13*). Extensive efforts have been dedicated to uncovering the regulatory mechanisms governing expression levels and interindividual variations within populations (*8*, *14*). To date, the quantitative description of expression strength and noise has been an important goal in biology to illustrate cellular behavior (*15*). However, this task has relied on fluorescence microscopy (*1*, *16*) and flow cytometry (FCM) (*15*, *17*) assays of individual clonal populations, which are time-consuming and labor-intensive when testing large amounts of genetic variants. Therefore, a general, precise, and high-throughput method for the profiling of expression properties is urgently needed.

To address the above issue, we focused on Sort-Seq (*18*–*21*) [also named FlowSeq and FACS-seq (fluorescence-activated cell sorting–sequencing)], by which each cell in a mixed library of cells with different expression intensities can be sorted into different bins based on different expression intensities and then quantified through next-generation sequencing (NGS) to derive the expression pattern of each genotype, including mean expression level and expression noise level. This approach has been broadly used in profiling sequence-function relationships associated with transcriptional regulation (*18*, *19*, *22*, *23*), translational regulation (*19*, *20*, *24*), regulatory RNAs (*25*, *26*), protein-sequence interactions (*27*), etc. In addition, the validity of Sort-Seq has been demonstrated in a wide range of organisms, including bacteria (*19*), yeast (*18*, *20*, *23*), and mammalian cells (*24*). However, it remains difficult to derive precise expression characteristics, especially the expression noise, from Sort-Seq data. Existing methods have focused on fitting the binned distribution to a log-normal (*20*, *21*, *25*, *28*) or gamma distribution (*18*, *22*, *29*). However, the representation capability of these probability densities may not always provide an exact fit (*6*, *30*, *31*). On the other hand, the parameter learning process of these methods still needs to be improved. For instance, apart from the binned distribution, other data, such as the overall fluorescence intensity density, should be considered. Hence, to obtain expression properties with high throughput and high precision, a common, rigorous data processing method for Sort-Seq is needed.

Therefore, we have developed dSort-Seq, a deep-learning–assisted Sort-Seq approach (Fig. 1). In this method, instead of using log-normal or gamma distribution, we applied a two-component log-Gaussian mixture model (LGMM) to match the steady-state gene expression density, which is more precise and robust in fitting the real data. To decode Sort-Seq data, we adopted a Bayesian neural network to perform parameter learning. These innovations substantially improve the accuracy of Sort-Seq to derive expression

[1]MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China. [2]Tsinghua-Peking Center for Life Sciences, School of Medicine, Tsinghua University, Beijing 100084, China. [3]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China. [4]Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. [5]Institute of Bioprocess and Biosystems Engineering, Hamburg University of Technology, Hamburg 21073, Germany. [6]Center of Synthetic Biology and Integrated Bioengineering, School of Engineering, Westlake University, Hangzhou 310024, China. [7]Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China.
*Corresponding author. Email: fhb_14@163.com (H.F.); chongzhang@tsinghua.edu.cn (C.Z.)
†These authors contributed equally to this work.

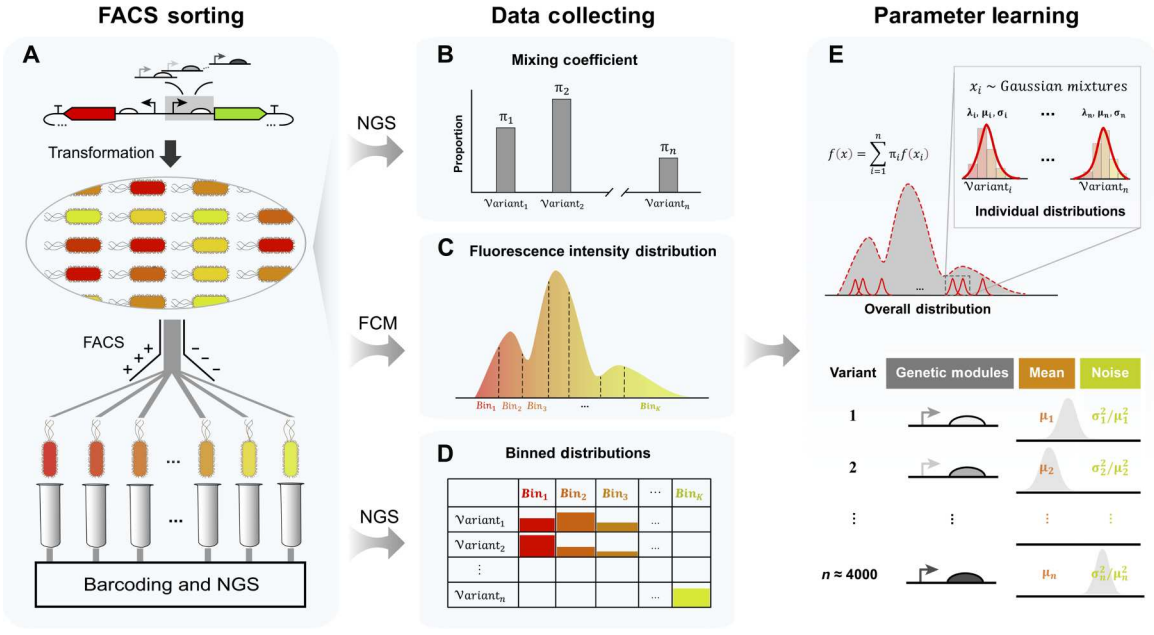

Fig. 1. Schematic overview of the dSort-Seq data workflow. (**A**) During Sort-Seq, a library with different expression patterns is sorted into customized bins based on the fluorescence intensity value. (**B**) The mixing coefficients are quantified via NGS. (**C**) The overall fluorescence density is measured by FCM, and the sorting boundaries are specified on the basis of the overall fluorescence intensity density. (**D**) The read count number across all bins as quantified by NGS reveals the binned distribution of each variant in the library. (**E**) Through parameter learning, the mean, expression noise, and their relationships can be precisely identified. μ, mean; σ, SD.

characteristics for thousands of variants. We demonstrated the validity of this pipeline from two aspects. First, dSort-Seq enables large-scale assays of dose-response relationships of biosensors with high precision, with which the desirable designs can be efficiently identified. Second, it also supports the high-throughput exploration of noise production mechanisms. For instance, we applied dSort-Seq to determine the effects of transcription and translation on expression noise in *Escherichia coli* and found them to have comparable contributions, contradicting the commonly accepted translational bursting mechanism (*3*). In addition, we also revealed that overlapping RpoD-binding sites would lead to high expression noise, which suggested an effective noise regulation strategy. Overall, our method, which provides considerable mathematical and biological insights, can serve as a promising high-throughput tool for use in various studies associated with gene expression.

RESULTS

DSort-Seq exhibits superior performance

Recent research on the stochastic nature of gene expression has shown that steady-state protein production in a clonal population follows a gamma (negative binomial) (*3*, *4*) or log-normal (*5*, *6*) distribution. However, neither of them can precisely match the real expression data (Fig. 2, A to C). To address this issue, dSort-Seq applied a two-component LGMM to represent the steady-state protein production density. This distribution was selected for several reasons, the first and foremost of which is that the mixture of Gaussians can theoretically approximate any continuous density given enough components (*32*), ensuring its ability to fit more complex densities compared with conventionally used models. In addition, the outliers [in more extreme cases, one peak of the bimodal expression densities (*31*)], which have a great impact on matching (*21*, *30*), can be viewed as being generated by a Gaussian component (*33*–*35*). To verify the model's ability to match expression distributions, we compared it with gamma and log-normal distributions in fitting quantitative datasets from independent resources (Fig. 2, A to C) (*28*, *36*); the quantile-quantile plots were applied to evaluate the fitting performances of different models (fig. S1). As a result, our method exhibited higher precision in representation. Hence, we used the LGMM for subsequent analyses.

Next, to derive gene expression characteristics from Sort-Seq data, we revisited the experimental procedure and considered incorporating more data into the parameter learning step (Fig. 1). To intuitively represent the dSort-Seq method, we defined the following terms: (i) the mixing coefficients, denoted by $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_n)$, where π_i is the proportion of the ith variant in the library; (ii) the log-scaled sorting boundaries, denoted by $\boldsymbol{b} = (b_0 = -\infty, b_1, ..., b_K = +\infty)$; (iii) the parameters involved in LGMM, denoted by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_n)$, where $\boldsymbol{\mu}_i = (\mu_{1i}, \mu_{2i})^T$, $\boldsymbol{\sigma}_i = (\sigma_{1i}, \sigma_{2i})^T$; and (iv) the probability of sorting the ith variant into the kth bin, denoted by P_{ik}. On the basis of these definitions, we built a Bayesian network to show the data generative process and dependencies among variables (see Materials and Methods; Fig. 2D). For parameter learning, instead of only matching the binned distribution via maximum likelihood estimation (MLE) as in previous methods (*21*, *22*, *28*), we constructed a Bayesian neural network to fit both the binned distribution and the overall fluorescence intensity density. Specifically, we designed two objective functions. The first was defined as the cross-entropy of the observed binned distribution relative to the theoretical binned distribution derived from LGMM (Fig. 2E), which, when minimized, optimizes the parameters of each LGMM to approximate the observed sorting data. The second objective was aimed

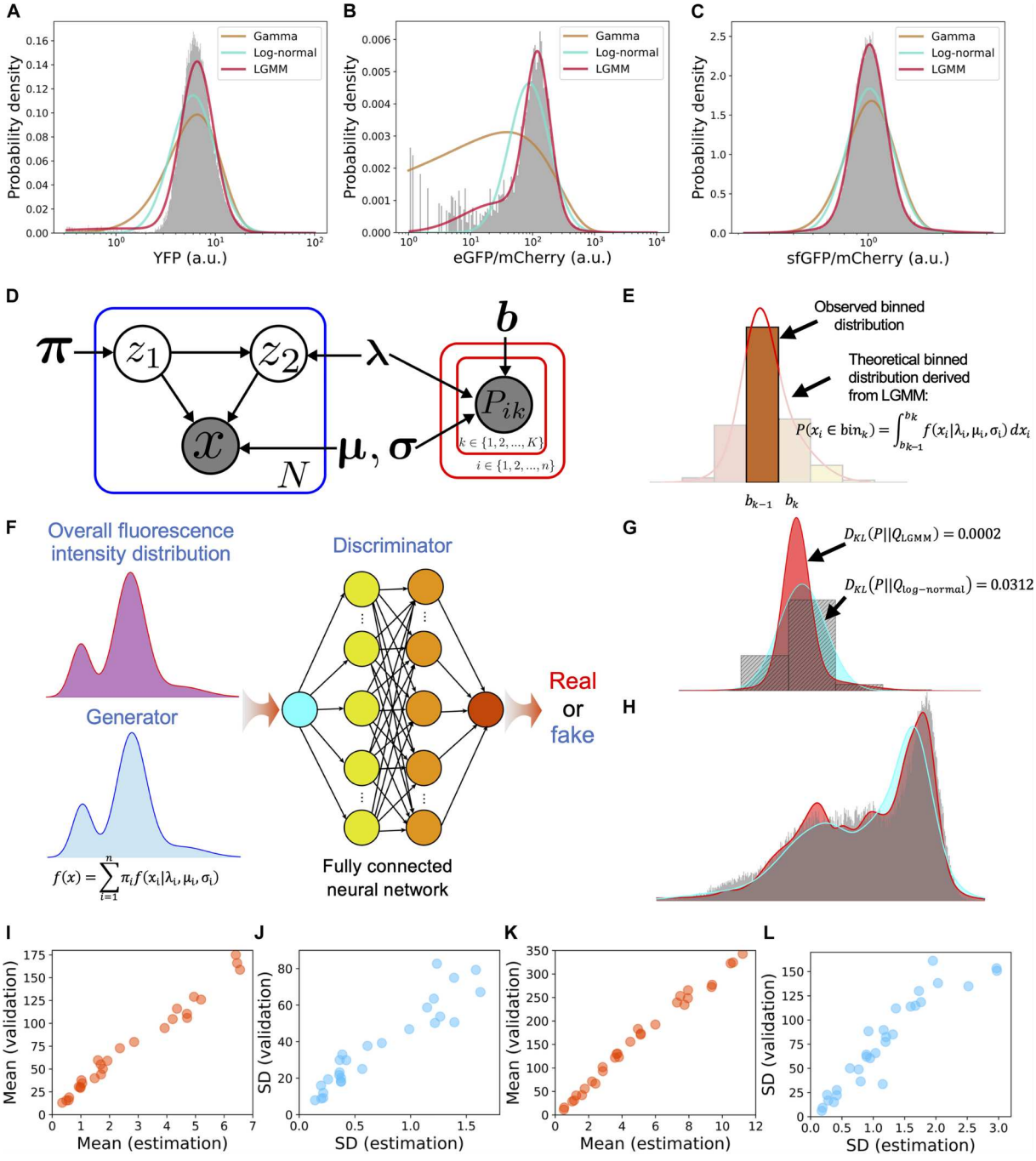

Fig. 2. Framework and performance of dSort-Seq. (**A** to **C**) Two-component log mixture of Gaussians better represents the gene expression distribution than conventional methods. (A) Gene expression controlled by the LmrA repressor (*36*). The histogram denotes the cytometry data of the unrepressed state. YFP, yellow fluorescent protein; a.u., arbitrary units. (B) Gene expression under the control of the *tnaC* variant K11R_CGC (*28*). The data were measured under 100 μM Ala-Trp. eGFP, enhanced green fluorescent protein. (C) Gene expression driven by the promoter *yebVp2* (this study). In (A) to (C), the gamma and log-normal distributions were matched using MLE, and the LGMM was fitted via the expectation-maximization algorithm. The red, cyan, and brown lines represent the fitting result of the two-component log mixture of Gaussian, log-normal, and gamma distributions, respectively. (**D**) Graphical representation of the model. (**E**) Theoretical fraction of the probability density within the corresponding boundaries. (**F**) Matching the mixture of two-component Gaussian mixture models to the overall fluorescence intensity distribution. The real data are sampled from experimental cytometry data; the fake data are generated from the LGMM. A fully connected neural network is used as a discriminator. (**G**) Example (V8A_GCC, 0 μM Ala-Trp, replicate 1) illustrating the superior performance of dSort-Seq in matching the binned distribution compared to the log-normal–based method. Kullback-Leibler divergence shows the performance of each fit. (**H**) Example (100 μM Ala-Trp, replicate 1) illustrating the superior performance of dSort-Seq in matching the overall fluorescence distribution compared to the log-normal–based method. In (G) and (H), the red and cyan distributions refer to the results derived from dSort-Seq and the log-normal–based method, respectively. The gray distribution refers to the real data. (**I** to **L**) Individually analyzed expression characteristics of reconstructed *tnaC* variants by cytometry highly correlated with those estimated via dSort-Seq in terms of means [(I) 0 μM Ala-Trp, $n = 26$; (K) 100 μM Ala-Trp, $n = 30$] and SDs [(J) 0 μM Ala-Trp; (L) 100 μM Ala-Trp).

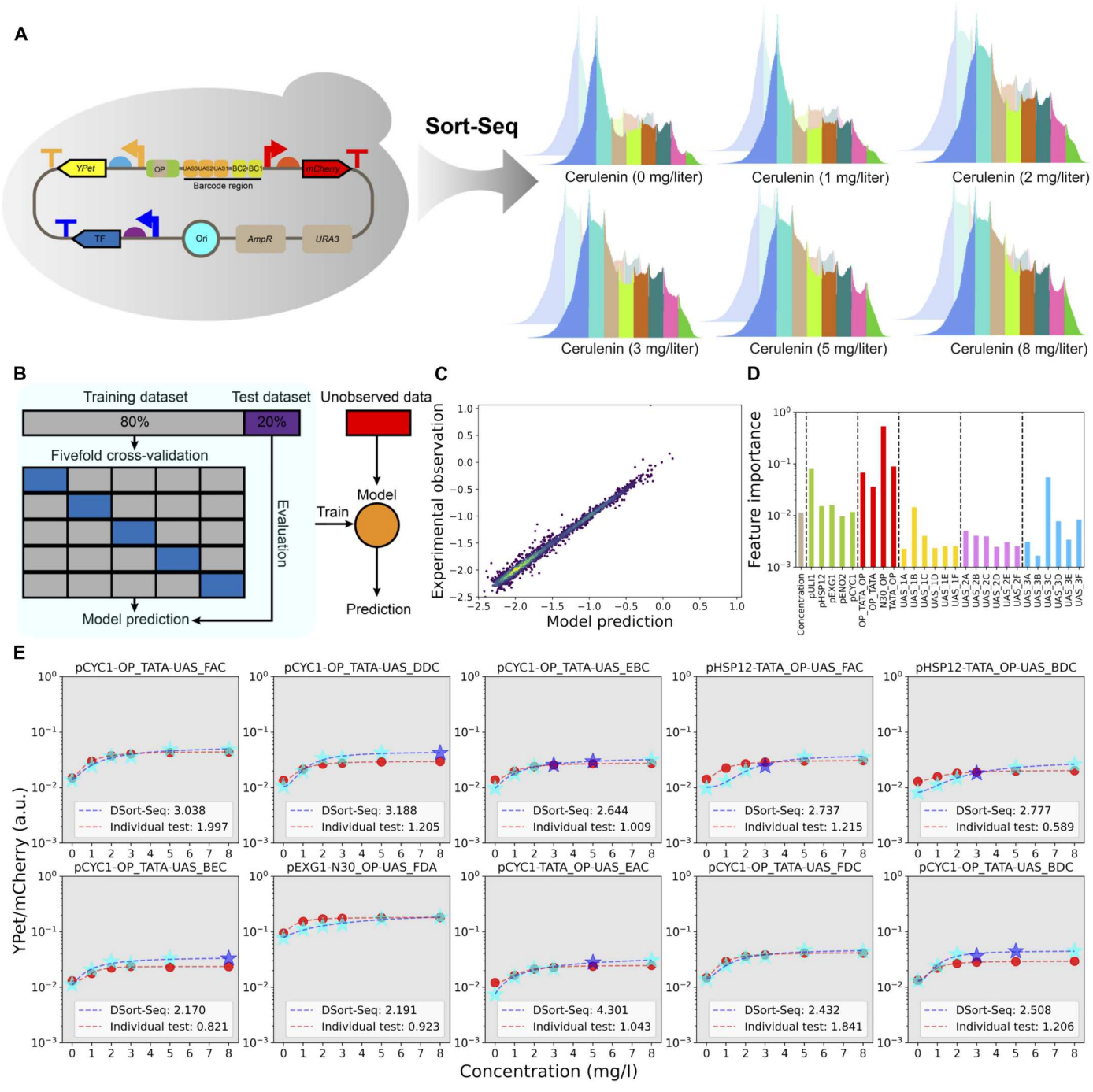

Fig. 3. The dSort-Seq profiling of FapR-*fapO*–based malonyl-CoA–dependent gene expression. (**A**) Sort-Seq characterization of the malonyl-CoA biosensor library under six different cerulenin concentrations (0, 1, 2, 3, 5, and 8 mg/liter). Cells were sorted into eight bins according to their responses to ligand. Two biological replicates were examined for each Sort-Seq experiment. (**B**) Schematic diagram of the machine learning process. Gradient boosting regression was used here to interpret the relationship between features and expression strengths. The hyperparameters were optimized through fivefold cross-validation; then, the whole training dataset was used to train the model parameters, and the test dataset was used to evaluate the generalization capacity of the model. Last, the model was trained on the entire observed dataset to obtain predictions for unobserved data. (**C**) The model performance in the test dataset showed a good generalization capacity (n = 3077). (**D**) Gini importance that contributes to the gradient boosting regression tree. (**E**) Dose-response curves of 10 combinations with substantial dynamic ranges. Data points represent the mean values of YPet/mCherry under different cerulenin concentrations, where red dots represent individual characterization data, cyan stars represent data from dSort-Seq characterizations, and blue stars denote data from machine learning predictions. The dashed lines represent response curves fitted by the Hill equation (see Materials and Methods). The dynamic ranges of dSort-Seq calculations and individual characterizations are labeled on each panel.

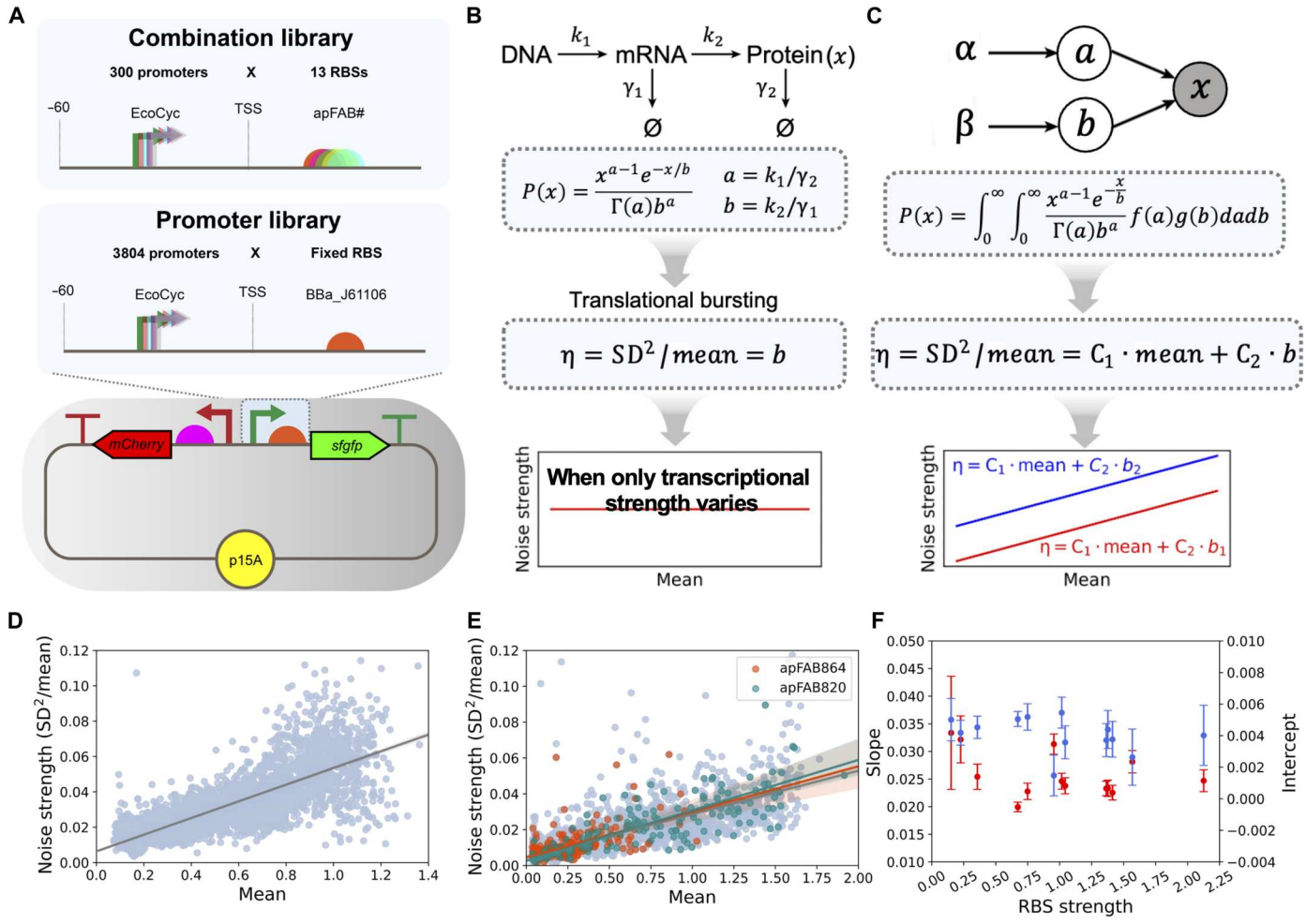

Fig. 4. The dSort-Seq profiling of transcriptional and translational effects on noise production in *E. coli* K12 MG1655. (**A**) Design schemes of the promoter and the combination libraries. TSS, transcriptional start site. (**B**) According to the translational bursting mechanism, steady-state protein production follows a gamma distribution (*3*). In this distribution, the parameter "*a*" represents the transcription rate and "*b*" represents the translation rate. As a corollary, the burst size, denoted by the Fano factor, is linearly correlated with the translation rate and independent of the transcription rate. (**C**) According to the hierarchical Bayesian model, the intercept in the relationship between noise strength and the mean expression level is proportional to translational strength, indicating that translational bursting still dominates noise production at low expression levels (*31*). In this figure, b_1 and b_2 ($b_1 < b_2$) denote the translation rate, and C_1 and C_2 are constants. (**D**) The noise strength is linearly correlated with the mean expression level when only transcriptional strength varies. The gray line exhibits the linear regression result, which is shaded to show the 95% confidence interval. (**E**) The relationships between noise strength and mean expression level are similar when the translation module varies. The gray, orange, and green lines represent the regressions of all combinations and combinations with RBS apFAB864 and apFAB820, respectively. [RBS strength: apFAB820 (1.57) > apFAB864 (0.36); see Materials and Methods]. (**F**) The linear regression slopes (red dots) and intercepts (blue dots), obtained through the least squares regression method, do not exhibit a positive correlation with RBS strength.

at matching the overall fluorescence intensity distribution of the whole library. For this purpose, we applied a generative adversarial network (Fig. 2F) (*37*). To elaborate, a generator was designed on the basis of the data generative process. A fully connected neural network was applied as the discriminator. During training, data generated from the generator are sent to the discriminator along with the real fluorescence intensity values, and then the discriminator determines whether each piece of data is real or not. Hence, a two-player game is played between the generator and the discriminator, and the overall fluorescence intensity distribution can be matched by the generator when they are in equilibrium. We included these two objectives in the Bayesian neural network, with which the parameters can be learned through backpropagation (fig. S2).

Subsequently, we tested the validity of dSort-Seq with data from our previous Sort-Seq profiling of a comprehensive codon-level mutagenesis library of *tnaC* (*28*). This experiment was performed under three different ligand concentrations (0, 100, and 500 μM Ala-Trp), each with two biological replicates (fig. S3A). However, by fitting each binned distribution to the nonrobust log-normal density, their results were obviously affected by outliers and could not precisely match the experimental observations in terms of both the binned distribution (Fig. 2G) and the overall fluorescence intensity distribution (Fig. 2H). In addition, the expression characteristics derived from the log-normal distribution were also subject to error (see Materials and Methods and figs. S4 and S5). Therefore, we applied dSort-Seq in this case to derive the expression properties (see Materials and Methods and data file S1). As a result, the strong

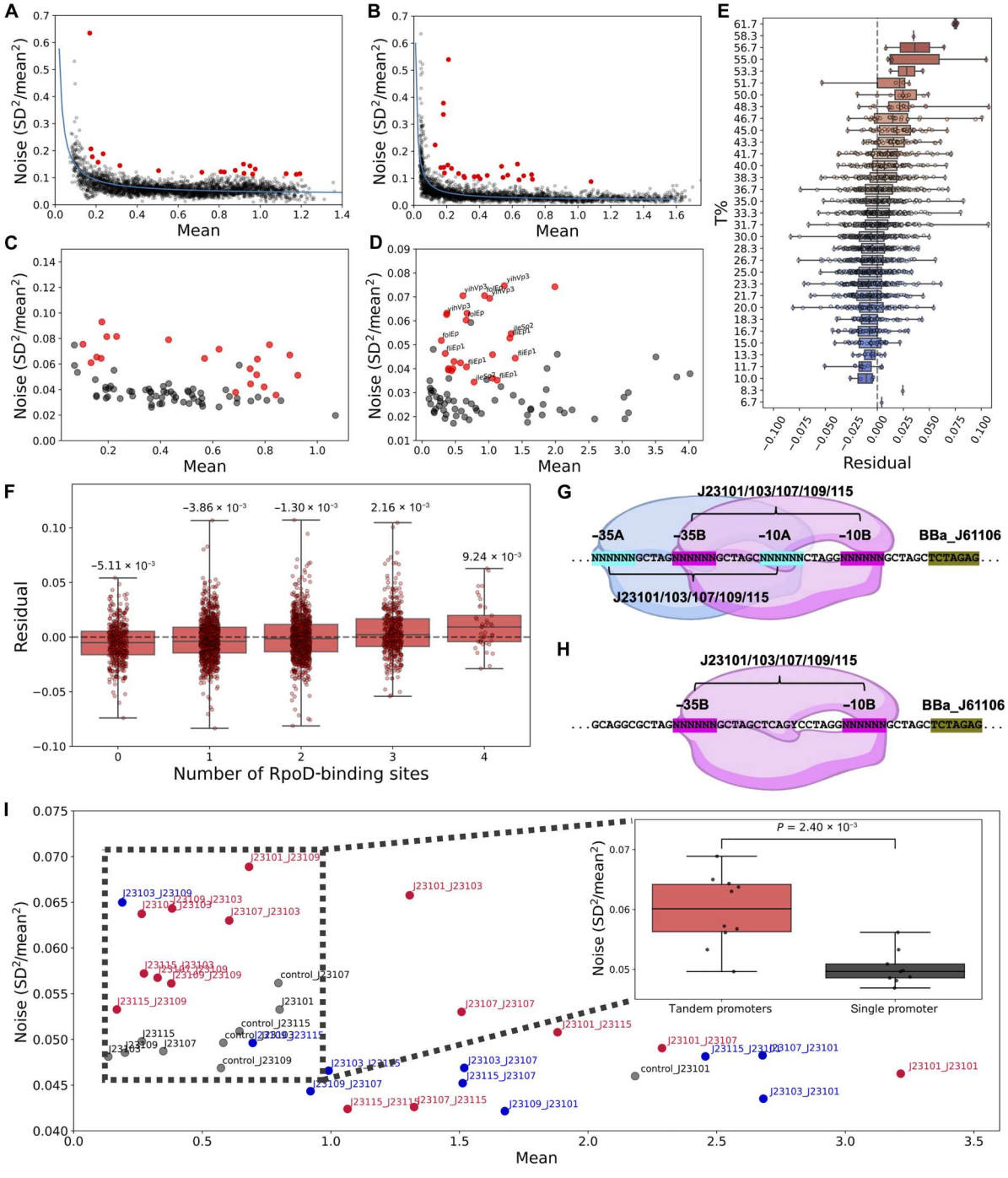

Fig. 5. Overlapping RpoD-binding sites result in high expression noise. (**A** and **B**) Correlation of the expression noise with expression strength in (A) promoter and (B) combination libraries. At low mean expression levels, the noise decreases as the expression strength increases; at high mean expression levels, the noise converges to a constant value. The blue lines show the regression results (see Materials and Methods). Twenty promoters and 25 combinations exhibiting high expression noise are marked as red dots. (**C**) Twenty sequences from the promoter library and (**D**) 25 sequences from the combination library showing high expression noise were constructed and assayed through FCM. Their expression noise (red dots) is higher than that of randomly selected variants (black dots) at their corresponding mean expression levels. (**E**) Promoters with higher levels of expression noise residuals were enriched in thymine. (**F**) Promoters with more RpoD-binding sites exhibited increased expression noise residuals. The respective box plots were annotated with the median residuals atop each group. (**G**) Design scheme of 25 tandem promoters, each containing two overlapping RpoD-binding sites. (**H**) Design scheme of five constitutive promoters with the same length as the tandem promoter, each with only one RpoD-binding site. (**I**) Compared to promoters with single RpoD-binding site, the tandem promoters exhibited significantly higher expression noise ($P = 2.40 \times 10^{-4}$, one-tailed t test), especially when the stronger promoter was positioned upstream of the weaker promoter. Red dots represent tandem promoters where the stronger promoter located upstream of the weaker promoter, blue dots represent tandem promoters where the stronger promoter located downstream of the weaker promoter, and gray dots denote promoters with a single RpoD-binding site. Data within the dashed box were included in the analysis because the promoters in this region had similar mean expression levels but exhibited varying levels of expression noise.

correlations of the mean (figs. S3, B to D; Pearson′s $r = 0.989$, 0.979, and 0.978 for 0, 100, and 500 μM Ala-Trp, respectively) and SD (figs. S3, E to G; Pearson′s $r = 0.914$, 0.903, and 0.891 for 0, 100, and 500 μM Ala-Trp, respectively) of expression between biological replicates indicated the reliability of dSort-Seq profiling. In addition, the individual validation data, even if measured via another flow cytometer, were highly consistent with the calculation results (Fig. 2I, mean for 0 μM Ala-Trp, Pearson′s $r = 0.991$; Fig. 2J, SD for 0 μM Ala-Trp, Pearson′s $r = 0.942$; Fig. 2K, mean for 100 μM Ala-Trp, Pearson′s $r = 0.994$; Fig. 2L, SD for 100 μM Ala-Trp, Pearson′s $r = 0.931$), which proved the model′s ability to precisely capture the expression characteristics.

DSort-Seq enables the screening of biosensors with desired response features

Given the superior performance of dSort-Seq in characterizing expression properties, we applied it to practical scenarios to highlight its applicability. First, as an example of expression strength mining, we focused on the metabolite biosensor, through which the intracellular concentration could be converted to a change in gene expression. The key performance indicators of a biosensor include sensitivity, specificity, dynamic range, and operational range (*38*), most of which can be determined from dose-response relationships. However, to the best of the authors′ knowledge, a method for large-scale profiling of the dose-response curves with high precision is lacking. Therefore, we applied dSort-Seq to address this problem. For instance, we tested it in our previously reported dataset by Zhou *et al.* (*39*), which contained Sort-Seq results for 5184 FapR-*fapO*–based biosensors, consisting of combinations of six transcription factor dosages (*pGPD*, *pENO2*, *pHSP12*, *pEXG1*, *pCYC1*, and *pULI1*), four operator insertion schemes (TATA_OP, OP_TATA, OP_TATA_OP, and N30_OP), and 216 arrangements of upstream enhancer sequences (UASs; 3 tandem UASs selected from UAS_A, UAS_B, UAS_C, UAS_D, UAS_E, and UAS_F). Each combination was encoded by a specific DNA barcode to ensure its identification via NGS. The library was transformed into *Saccharomyces cerevisiae* BY4700 and assayed through Sort-Seq under six different cerulenin concentrations (0, 1, 2, 3, 5, and 8 mg/liter), each with two biological replicates (Fig. 3A). However, owing to the limited precision and robustness of log-normal–based analysis, the dose-response curves derived from Sort-Seq were susceptible to error. Therefore, we applied dSort-Seq to this case to determine whether it could accurately evaluate the response performance. The resulting responses showed strong correlations between biological replicates at different concentrations (fig. S7; Pearson′s $r > 0.99$ for all experimental conditions), indicating the reliability of the calculation.

As the library was nonuniform, we could obtain only 16,960 expression strengths of 2960 combinations through dSort-Seq. To obtain the rest of the data, we applied a machine learning approach to generate predictions. Specifically, we used one-hot encoding to transform each biosensor combination into a 27-dimensional vector. This included 5 dimensions for the promoters of the transcription factor, 4 dimensions for the operator insertion schemes, and 18 dimensions for the sequences of the tandem UAS (UAS_1A/B/C/D/E/F, UAS_2A/B/C/D/E/F, and UAS_3A/B/C/D/E/F). These vectors served as input features along with the cerulenin concentration. Gradient boosting regression was applied to fit the log-scaled expression strength (see Materials and Methods). Note that as combinations containing the promoter *pGPD* suffered from a heavy metabolic burden, leading to them being underrepresented in the library, we excluded them from the machine learning analysis. Therefore, the data used to train the model contained 15,382 responses, covering 59.2% of the whole combinatorial space (data file S2). To avoid overfitting, we randomly split the dataset into two subgroups, with 80% of the data used as the training dataset to optimize the hyperparameters through fivefold validation as well as train the model parameters. The remaining 20% were used as the test dataset to check the generalization capacity of the model (Fig. 3B). The performances in the test dataset (Fig. 3C; $r^2 = 0.988$) indicated that the model had reasonable generalization capacity and captured biological signals. Subsequently, we trained the model on the whole dataset and predicted the uncharacterized responses. The dSort-Seq data accompanied by the predicted results were then applied to generate the dose-response curves for all combinations (data file S2). We validated these dose-response relationships using 92 individual characterization results that were previously obtained (*39*). Linear regression was applied to fit the data values within the same scale. The resulting high consistencies (Pearson′s $r > 0.96$ for all cases; fig. S8) demonstrated that with dSort-Seq and machine learning, the expression properties of the enormous combinatorial space could be effectively explored. In addition, these data demonstrated a clear decrease in the precision of response values calculated using the log-normal fitting method compared to dSort-Seq (fig. S9).

We then analyzed the features that contributed most to model predictions via Gini importance (Fig. 3D). Overall, consistent with previous discoveries (*39*), the responses of the biosensor were mostly affected by the operator insertion schemes. In addition, a strong determinant of the result was observed if the third UAS was UAS_C. Next, we focused on optimizing the dynamic range of the biosensor, which measures its signal-to-noise ratio. To this end, we fitted each dose-response relationship to the Hill equation (Eq. 17) and derived the corresponding dynamic range using the obtained parameters (Eq. 18; see Materials and Methods). Ten combinations with obvious dynamic ranges (top 60) were randomly selected, constructed, and assayed through FCM (Fig. 3E). Their response performances were consistent with dSort-Seq and machine learning calculations, with pCYC1-OP_TATA-UAS_FAC achieving the highest dynamic range of 2.0. It is important to note that the dynamic ranges obtained from library characterization and individual characterization did not perfectly match, likely because of slight variations in media formulation (see Materials and Methods) and experimental fluctuations. For instance, the dynamic range of pHSP12-TATA_OP-UAS_FAC is 1.22 in Fig. 3E and 3.83 in fig. S10. Nevertheless, dSort-Seq remained useful in identifying biosensors with substantial responses, as most combinations in the library do not interact with cerulenin (fig. S10). In addition, UAS_C was the third UAS in most of the top 60 sequences with the highest dynamic ranges (83.3%; data file S2), indicating its importance for interactions between yeast synthetic promoters and FapR when located at the third position. In addition to dynamic range, other indicators, including operational range and sensitivity, can also be evaluated and optimized in a similar manner. Therefore, with dSort-Seq, the designs with desired response features can be effectively identified.

DSort-Seq profiles the mean-noise landscape of *E. coli* endogenous promoters

In addition to the expression strength, expression noise is also an important factor affecting gene expression that leads to phenotypic diversity among genetically identical individuals. Previous association studies have found that expression noise is a heritable trait (*40*) and is determined by expression modules (*15*–*17*, *31*). Hence, for a given organism, how different expression modules shape the patterns of noise is a fundamental question. On the other hand, in terms of noise production mechanisms, the commonly accepted translational bursting model suggests that the protein within a cell is produced in bursts, where the burst size (noise strength, $\eta = SD^2/\text{mean}$) is related to only translation and is independent of transcription (Fig. 4B) (*3*, *4*). However, relevant experimental evidence is still scarce. The problem with the translational bursting mechanism is not only that the gamma distribution cannot accurately represent the gene expression distribution (*6*) but also that it cannot interpret the dependence between transcription and noise strength (*16*). However, limited by the low throughput of classic quantitative methods, research on transcriptional and translational contributions to expression noise is always based on the analysis of a small amount of data (*16*, *41*), which is susceptible to experimental error as well as outlier samples. To address these issues, our approach may serve as a promising method because of its ability to produce high-quality data in a massively parallel manner. Therefore, as a proof of concept, we performed systematic profiling of transcriptional effects on expression noise in *E. coli* based on dSort-Seq.

For library construction, we collected 3804 endogenous promoters of *E. coli* K12 MG1655 from the EcoCyc database (*42*) (https://ecocyc.org/), for which the 60-nucleotide (nt) sequence upstream of each transcription start site was regarded as the promoter region. The oligonucleotide library composed of collected promoters was high-throughput synthesized and assembled into a low–copy-number, dual-fluorescence plasmid, pMPTPV_dual_fluorescence, in which a superfolder green fluorescent protein (sfGFP) was used as the response reporter that was under the control of a particular promoter with a fixed ribosome-binding site (RBS) BBa_J61106. In addition, a constitutively expressed reporter, mCherry, served as an internal reference to eliminate cell-to-cell variations such as cell volume and plasmid copy number (Fig. 4A and fig. S14). It is important to note that, as the expression cassette of mCherry was fixed, its intrinsic noise did not affect the results and could be disregarded. After electroporation into MG1655 cells, a cell library with broad levels of sfGFP expression was obtained. We characterized the cultivated cell library by FCM with three biological replicates and then sorted into 12 bins based on the fluorescence intensity of sfGFP relative to mCherry, followed by NGS to quantify the proportion of each variant in each bin. The acquired datasets were then processed and analyzed by our method. The results showed that 2920 (76.8% of the total library) promoters were highly consistent among all three replicates (fig. S18 and data file S3). To validate the result, 60 single colonies with different genotypes were randomly selected for individual cytometry assays. They were cultured and measured using the same conditions and settings as the Sort-Seq experiment (fig. S19). The strong consistency with the dSort-Seq results (fig. S20; Pearson's r = 0.981 and 0.921 for the mean and SD, respectively) indicated the reliability of the profiling. Autofluorescence was quantified by assaying the pMPTPV strain with only mCherry expression and no sfGFP expression, and the result showed that autofluorescence could be neglected relative to the fluorescence intensity of each candidate of the library (fig. S13).

The bulk data generated a comprehensive landscape of promoter strength and expression noise along the *E. coli* genome [data file S3; we also visualized it through D3GB (*43*) at http://thu-big.net/Escherichia_coli_K12_MG1655_promoters], which was beneficial for understanding the transcriptional strategies for different genes. For instance, we investigated whether essential and nonessential genes of *E. coli* have different expression patterns (see Materials and Methods). As a result, the essential genes showed greater transcriptional intensities than the nonessential genes (fig. S22; $P = 4.22 \times 10^{-16}$, one-tailed t test). Given that the high transcriptional strength is usually related to low expression noise (*16*), these functionally important genes are more likely to confer lower levels of noise, which is consistent with the results of a previous genome-wide association study (*17*). However, it should be noted that there are many reasons why essential genes may be expressed at higher levels, including the frequent requirement for important genes at high levels. Subsequently, we investigated the relationship between noise strength and mean expression level. The results showed that the noise strength was linearly correlated with the expression strength when the transcription module varied (Fig. 4D; Pearson's r = 0.745); hence, transcription contributed to expression noise. This discovery, however, is inconsistent with the inference of the translational bursting model (Fig. 4B), suggesting the limitation of the model in interpreting noise production mechanisms in *E. coli*.

Transcription and translation make comparable contributions to noise production

Although the translational bursting mechanism is unable to account for the contribution of transcription to noise production, the hierarchical Bayesian model, developed by introducing transcriptional and translational fluctuations into the translational bursting model, can successfully explain this phenomenon (*31*). However, the model showed that different translation modules would lead to varying intercepts (*16*, *31*) in the relationship between noise strength and the mean expression level (Fig. 4C, $\eta = C_1 \cdot \text{Mean} + C_2 \cdot b$, where b represents the translation strength and C_1 and C_2 are constants). Hence, at low levels of transcription, the expression noise is still dominated by translation and can be controlled independently of mean protein abundance by combining strong transcription modules with weak translation modules for low noise or weak transcription modules with strong translation modules for high noise. Although this conclusion has been supported by a fluorescence microscopy experiment analyzing 40 *Bacillus subtilis* strains expressing GFPmut3 with different combinations of transcription and translation modules (*16*), further verification is needed, particularly in Gram-negative strains due to the varying expression modes among different types of strains (*44*, *45*). Furthermore, the contribution of transcription and translation to noise production has not been comprehensively and directly observed thus far. As our method has greatly expanded the test throughput of expression noise, we applied dSort-Seq here to examine these features.

Therefore, we designed a combination library comprising different combinations of 300 promoters and 13 RBSs. The promoters were randomly selected from the EcoCyc database; whereas the RBSs, which were chosen for their varying translational strengths,

were from the BIOFAB reporter plasmid series (Fig. 4A) (*46*). We prepared the combination library in the same manner as the promoter library in the pMPTPV_dual_fluorescence plasmid. After electroporation, we performed a dSort-Seq assay of the cell library, with three independent biological replicates to ensure the reliability of the results. After data processing, 2733 combinations (70.1% of the whole library) were highly consistent among the replicates and were retained for subsequent analysis (fig. S24 and data file S4). Subsequently, 60 single colonies with different genotypes were randomly picked and assayed individually with FCM (fig. S25). Their means and SDs of expression were strongly correlated with the dSort-seq results (fig. S26; Pearson's $r = 0.976$ and 0.937 for the mean and SD, respectively), proving the validity of the profiling.

We then performed regression analysis between noise strength and the expression mean for different translational modules (Fig. 4E and fig. S28) to test the hierarchical Bayesian model. However, their correlation barely changed when the translation module varied in terms of both slope and intercept (Fig. 4F and table S6). Instead, our results showed that the noise strength was highly coupled with the mean expression level, indicating the difficulty of adjusting expression noise independently of the mean protein abundance by tuning the strength of the transcription and translation modules. In other words, if altering the transcription and translation modules produces similar effects on the mean expression level, then their effects on noise strength should also be similar. Furthermore, to determine whether translation bursting dominates noise production at low transcription levels, we constructed several weak expression combinations and performed cytometry assays. As a result, the combinations with comparable mean expression levels showed similar fluorescence intensity distributions (fig. S29), suggesting that the contributions of transcription and translation to noise are comparable, even in the case of weak expression strength. It is worth mentioning that the discrepancy of our results with previous fluorescence microscopy observations (*16*) may stem from the different dependence between transcription and translation in *E. coli* and *B. subtilis* (*44*, *45*), which is not accounted for in the translational bursting mechanism or the hierarchical Bayesian model.

Overlapping RpoD-binding sites can lead to high expression noise

We then analyzed the relationship between expression noise ($CV^2 = SD^2/Mean^2$) and strength (Fig. 5, A and B). The expression noise exhibited a strong negative correlation with the mean protein abundance at low levels of expression and then reached a plateau after a critical point, which is consistent with previous observations (*16*, *31*). In addition to the general correlation, some unique expression features also piqued our interest, especially for those sequences exhibiting high expression noise at their corresponding mean expression levels. To ensure that these outliers were not the results of experimental error, we reconstructed and assayed 20 high-noise candidates from the promoter library and 25 from the combination library and then performed an individual cytometry assay (see Materials and Methods and figs. S30 and S32). The credibility of the discovery was demonstrated by the high consistency between the expression characteristics obtained from dSort-Seq calculations and individual characterizations (figs. S31 and S33; Pearson's $r > 0.95$ for the promoter library and > 0.84 for the combination library). Moreover, the expression noise of these variants was apparently higher than that of randomly selected colonies (Fig. 5, C and D). Notably, among the 25 high-noise combinations, several promoters appeared frequently (e.g., *fliEp1*, *ileSp2*, *yihVp3*, and *folEp*; Fig. 5D), suggesting that the extra noise may be derived from transcription rather than translation.

Next, to identify the factor that contributed to the additional expression noise, we sorted the promoter sequences based on the regression residuals of expression noise (see Materials and Methods). We found a positive correlation between thymidine proportion and residuals (Fig. 5E). Furthermore, there was no common regulator was found to be associated with the sequences within the same group (data file S5). Therefore, we focused on transcription initiation factors, especially RpoD (σ^{70} factor), which transcribes most genes in *E. coli*. Promoters recognized by RpoD generally contain two consensus hexamers centered at 10 and 35 nt upstream of the transcription start site. These two regions are rich in adenosine and thymine, especially thymine [average of 4.73 per promoter compared to 3.75 for adenosine, 1.80 for cytosine, and 1.70 for guanine (*47*)]. On the basis of this, we hypothesized that the high-noise promoters contain more RpoD-binding sites. To test this hypothesis, we searched the DPinteract database (*48*) for potential RpoD-binding sites for each promoter (see Materials and Methods). The analysis revealed that promoters with a greater number of RpoD-binding sites exhibit increased expression noise (Fig. 5F), indicating a potential association between these factors.

Subsequently, to determine whether overlapping RpoD-binding sites would result in high expression noise, we constructed 25 tandem promoters based on combinations of 5 Anderson promoters (J23101/103/107/109/115; Fig. 5G) to drive the expression of *sfgfp*. In addition, the five constitutive promoters were also individually constructed as controls. To exclude the effect of promoter length and the −35 region-proximal sequence on the results, we also constructed five promoters of the same length as the tandem promoters, while preserving only the downstream RpoD-binding site (Fig. 5H). Subsequently, we performed an individual cytometry assay of the 35 promoters (fig. S35). As a result, the gene expression driven by tandem promoters showed higher noise compared to a single promoter (Fig. 5I; $P = 2.4 \times 10^{-4}$, one-tailed t test), especially when the stronger promoter was positioned upstream of the weaker promoter (Fig. 5I), suggesting that the transcriptional interference caused by the occlusion of promoters contributed to noise production. Hence, our results uncovered a feasible and simple noise modulation strategy in *E. coli* by tuning the number and relative positions of sigma factors upstream of the transcription start site.

DISCUSSION

Gene expression dosage is directly associated with a variety of phenotypes of a population (*49*); hence, there is no doubt that high-throughput profiling will deepen our understanding of cellular behavior. Here, we focused on biosensors that can sense metabolite concentrations and regulate gene expression. According to the different output signals, biosensors have various applications, including in high-throughput screening (*50*), medical diagnosis (*51*), and cell imaging (*52*). For each application, the dose-response relationship is a key indicator that needs to be tuned to meet practical needs. Fortunately, dSort-Seq was shown to be a powerful tool for characterizing and optimizing the biosensor response performance in a high-throughput and high-precision manner, enabling the

engineering of biosensors with desired properties. Compared to positive and negative screening (*53*, *54*), by which only the dynamic range could be optimized, dSort-Seq can yield more comprehensive information to meet the needs of various situations. In addition, it is much more efficient than traditional trial-and-error approaches. Therefore, dSort-Seq provides a solution for profiling the expression landscape of the combinatorial sequence space. On the other hand, the high-quality dSort-Seq dataset also has the potential to serve as a basis for deciphering physiological mechanisms (*26*, *28*).

Noise in biological systems has been widely demonstrated to influence various intracellular processes (*55*) and the physiological properties of a population (*13*, *56*), although the effects of noise strength vary across different situations. For instance, low noise can ensure stable biosynthesis pathways and robust synthetic gene circuits (*57*). In contrast, high phenotypic variability promotes evolvability (*58*–*60*). Therefore, it is necessary to understand the origin of the noise, as well as control noise rationally for various applications. Regarding noise regulation, various strategies have been proposed to control gene expression noise independently, including engineering transcription and translation in synthetic gene circuits (*61*, *62*), introducing pulsatile input to control the promoter activation frequency and transcription rate independently (*63*) and expressing two copies of the target gene from separate circuits with different characteristics (*64*), among others. Through dSort-Seq profiling of different combinations of promoters and RBSs in *E. coli*, the transcriptional effect on gene expression noise was revealed. Specifically, a higher thymidine proportion without position preference in the promoter sequence would lead to a higher level of noise. One hypothesis to explain this phenomenon is that RpoD-binding sites, which are rich in thymine, influence noise production. We have demonstrated that promoters with overlapping RpoD-binding sites contribute to noise production due to occlusion of promoters. However, a detailed molecular biological explanation for this phenomenon is still lacking. It is worth mentioning that in *E. coli*, 831 genes have been found to be under the control of tandem promoters (*65*), suggesting the broadness of the regulatory scheme. Hence, in-depth research on modeling molecular events in the transcription process is needed to elucidate the effect of promoter architecture on expression noise.

From the methodology point of view, the design-build-test-learn cycle is emerging as a key workflow in synthetic biology, where the test is the rate-limiting step due to its low throughput (*66*). The development of Sort-Seq has undoubtedly greatly extended the test throughput, enabling more efficient characterization and optimization of biological parts. The dSort-Seq approach shows that the learning can be encapsulated into the test to improve its capability by modeling the data generative process for the high-throughput experiment. Moreover, it is worth mentioning that as the ability of the Gaussian mixture models in distribution matching can be improved by increasing the number of mixture components, dSort-Seq can be easily transferred to more complex situations in which multiple feedback circuits are involved (*67*). Thus, this pipeline has great potential to determine the mean-noise space for various gene expression modules, providing diverse synthetic parts that can be applied to different fields, such as biosynthesis (*68*), laboratory-based adaptive evolution (*59*), transcriptional regulation (*69*), and protein-protein interactions (*70*, *71*). Overall, owing to the flexibility, high precision, and high throughput of this method, we believe dSort-Seq can serve as a powerful tool that provides a wide range of novel research opportunities.

MATERIALS AND METHODS

Parameter-learning algorithm of dSort-Seq

We represent the log-scaled expression density of each variant by a two-component Gaussian mixture model (where x_i denotes the log-scaled intensity value of the ith variant)

$$x_i \sim f(x_i|\lambda_i,\mu_i,\sigma_i) = \lambda_i N(\boldsymbol{\mu}_{1i},\boldsymbol{\sigma}_{1i}^2) + (1-\lambda_i)N(\mu_{2i},\sigma_{2i}^2)\ (i = 1,2,\ldots,n) \quad (1)$$

Hence, the overall logarithmic fluorescence intensity distribution can be modeled as a mixture of Gaussian mixture models

$$x \sim f(x) = \sum_{i=1}^{n}\pi_i f(x_i) \quad (2)$$

The generative process for each fluorescence intensity value should be as follows: (i) Choose a variant z_1~categorical($\boldsymbol{\pi}$), (ii) choose a Gaussian component z_2 from Bernoulli($z_2|z_1$), and (iii) choose a log-scaled intensity value x from $N(x|z_1, z_2)$ (Fig. 2D). The distributions of variables involved in the model are listed as follows

$$P(z_1 = i) = \pi_i\ (i = 1,2,\ldots,n) \quad (3)$$

$$P(z_2 = 1|z_1 = i) = \lambda_i \quad (4)$$

$$P(z_2 = 0|z_1 = i) = 1-\lambda_i \quad (5)$$

$$f(x_i|z_1 = i, z_2 = 1) = N(\mu_{1i},\sigma_{1i}^2) \quad (6)$$

$$f(x_i|z_1 = i, z_2 = 0) = N(\mu_{2i},\sigma_{2i}^2) \quad (7)$$

$$P_{ik\_\text{theoretical}} = \int_{b_{k-1}}^{b_k}\lambda_i N(x_i|\mu_{1i},\sigma_{1i}^2) + (1-\lambda_i)N(x_i|\mu_{2i},\sigma_{2i}^2)dx_i\ (k = 1,2,\ldots,K) \quad (8)$$

Among the parameters involved in the model, the sets $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ cannot be identified experimentally. To estimate them, we designed a probabilistic artificial neural network in which a double-objective optimization is performed. The first objective function is defined as the cross-entropy (H) of the observed binned distribution relative to the integral of the probability density over adjacent boundaries (Fig. 2E), which is shown as follows

$$\text{Min } H = -\sum_{i=1}^{n}\sum_{k=1}^{K}\{-P_{ik}\log[\int_{b_{k-1}}^{b_k} f(x_i|\lambda_i,\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i)dx_i]\} \quad (9)$$

By minimizing the above loss function, the binned distribution can be fitted. The other objective is to match the overall fluorescence intensity density. To this end, a generative adversarial network (*37*) is applied. Specifically, a generator is constructed on the basis of the abovementioned generative process. For the discriminator, a fully

connected neural network is used to determine whether the data are real or fake (Fig. 2F). During training, a two-player game is played between the generator G and the discriminator D with value function $V(G, D)$

$$\min_G \max_D V(G,D) = E_{x_{\text{true}}}\{\log[D(x_{\text{true}})]\} + E_{x_{\text{fake}}\sim f(x|\boldsymbol{\pi},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\sigma})}\{\log[1-D(x_{\text{fake}})]\} \quad (10)$$

Combining the above two parts, we can obtain the whole algorithm for parameter learning, as shown in fig. S2.

Obtaining expression characteristics from cytometry data through LGMM fitting

For each cytometry assay, the $\log_{10}$-transformed fluorescence intensity distribution was fitted by a two-component Gaussian mixture model (figs. S4, S5, S13, S19, S25, S29, S30, S32, and S35) via the expectation-maximization algorithm, which resulted in a representation of $f(x_i) = \lambda_i N(\mu_{1i}, \sigma_{1i}^2) + (1-\lambda_i)N(\mu_{2i}, \sigma_{2i}^2)$. The mean expression strength was calculated with Eq. 11

$$\begin{aligned}\text{Mean} &= \text{Mean}_1 \times \text{Mean}_2 \\ &= \exp(m_1 + V_1/2) \times \exp(m_2 + V_2/2)\end{aligned} \quad (11)$$

where $m_1 = \lambda_i\mu_{1i}\log(10)$, $V_1 = [\lambda_i\sigma_{1i}\log(10)]^2$, $m_2 = (1-\lambda_i)\mu_{2i}\log(10)$, and $V_2 = [(1-\lambda_i)\sigma_{2i}\log(10)]^2$. The SD of each expression density was calculated with Eq. 12.

$$\text{SD} = \sqrt{\text{Var}_1 \times \text{Var}_2 + \text{Var}_1 \times \text{Mean}_2^2 + \text{Var}_2 \times \text{Mean}_1^2} \quad (12)$$

where $\text{Var}_1 = \exp(2m_1 + V_1)[\exp(V_1 - 1)]$ and $\text{Var}_2 = \exp(2m_2 + V_2)[\exp(V_2 - 1)]$.

Comparison of dSort-Seq and the log-normal–based method

The dSort-Seq results were calculated as mentioned above, and the log-normal results were obtained from previously reported data (*28*) (note that because the actual slope of the sorting boundary lines on the log-log plot of enhanced GFP–mCherry in the experiment was 0.8810 instead of 1, each boundary value was shifted to the right by 0.3801 compared to the previous analysis; table S4). Next, as an example, we compared the performances of the two methods in matching the binned distribution of variant V8A_GCC under 0 μM Ala-Trp and calculated the Kullback-Leibler divergences (Eq. 13) of the observation from the theoretical binned distributions derived from the log-normal–based method and dSort-Seq. The results showed that dSort-Seq is more precise and robust than log-normal (Fig. 2G)

$$D_{KL}(P||Q) = -\sum\nolimits_x P(x)\log\left[\frac{Q(x)}{P(x)}\right] \quad (13)$$

Subsequently, we compared these two methods in fitting the overall fluorescence intensity distribution. For instance, we calculated the theoretical log-scaled overall distribution (100 μM Ala-Trp, replicate 1) derived from the log-normal–based method (Eq. 14) and dSort-Seq (Eq. 2). As a result, dSort-Seq also showed better performance (Fig. 2H)

$$f_{\text{log-normal}}(x) = \sum_{i=1}^{n}\pi_i N(x_i|\mu_i, \sigma_i^2) \quad (14)$$

Moreover, as mentioned above, the log-normal–based method is unable to fit the individual cytometry data, which usually serve as criteria for validating the Sort-Seq results (figs. S3 and S4). To measure the error, we calculated the expression strength and SD of individual validation data with both log-normal (Eqs. 15 and 16) and the LGMM (Eqs. 11 and 12), where the LGMM results were applied as ground truth to evaluate the precision of log-normal results. As a result, the response and SD inferred from log-normal showed notable deviations (fig. S6). Therefore, with log-normal, it is difficult to infer accurate expression properties from Sort-Seq experiments

$$\text{Mean}_{\text{log-normal}(\mu_i,\sigma_i)} = \exp\{\log(10)\cdot\mu_i + [\log(10)\cdot\sigma_i]^2/2\} \quad (15)$$

$$\text{SD}_{\text{log-normal}(\mu_i,\sigma_i)} = \sqrt{\{\exp[\sigma_i\log(10)]^2 - 1\}\ \exp\{2\mu_i\log(10) + [\sigma_i\log(10)]^2\}} \quad (16)$$

Strains and growth media

Molecular cloning was performed using *E. coli* DH5α (BioMed) as the host. *S. cerevisiae* BY4700 (MATa ura3Δ0) was obtained as a gift from DaiLab (J. Dai). *E. coli* K-12 MG1655 was from the American Type Culture Collection 700926. The Luria-Bertani (LB) broth contained yeast extract (5 g/liter; Oxoid), NaCl (10 g/liter; Sinopharm), and tryptone (10 g/liter; Oxoid). The synthetic complete medium minus Ura (SC-Ura) contained glucose (20 g/liter; Solarbio), yeast nitrogen base without amino acids (6.7 g/liter; Solarbio), and yeast synthetic drop-out medium supplements without uracil (2 g/liter; Sigma-Aldrich). Note that the composition of drop-out supplements used in our study was slightly different from that used by Zhou *et al.* (*39*, *72*), specifically with a higher concentration of *p*-aminobenzoic acid. The yeast extract-peptone-dextrose (YPD) medium contained YPD broth powder (50 g/liter; Solarbio).

DNA manipulation and reagents

Plasmid extraction and DNA fragment purification were performed using the Plasmid DNA Mini Kit I and the Gel Extraction Kit from Omega Bio-Tek, respectively. Polymerase chain reactions (PCRs) were carried out using the KAPA HiFi PCR Kit from KAPA Biosystems [95°C for 3 min, 25 cycles (98°C for 20 s, 63°C for 15 s, and 72°C for 5 s), 72°C for 30 s]. The restriction enzyme FastDigest Esp 3I (namely, Bsm BI) and T4 DNA ligase were purchased from Thermo Fisher Scientific. Cerulenin was ordered from APExBIO. All strains and plasmids used in this work are summarized in table S1. All oligonucleotides (table S2) were ordered from Azenta. The concentrations of the antibiotics kanamycin and ampicillin were 50 and 100 mg/liter, respectively. In all experiments, bacteria and yeast were grown at 37° and 30°C, respectively.

Gradient boosting regression

Gradient boosting regression was applied to predict the log-scaled expression strength. During training, the hyperparameters were optimized following the given order (min_samples_split, max_depth, min_samples_leaf, max_features, subsample, learning_rate, and n_estimators) through the grid search method. The optimized hyperparameters were min_samples_split = 5, max_depth = 8, min_samples_leaf = 1, max_features = 6, subsample = 0.95, learning_rate = 0.05, and n_estimators = 2800.

Fitting the dose-response relationship to the Hill equation

Each dose-response relationship was fitted by Eq. 17 via the Levenberg-Marquardt algorithm

$$S = S_0 + \frac{S_m - S_0}{1 + (C_{1/2}/C)^h} \quad (17)$$

where S_0 and S_m are the values of the sensor response at zero and saturating ligand concentrations, $C_{1/2}$ is the concentration at half saturation, and h is the Hill coefficient. The lower bounds and upper bounds of (S_0, S_m, $C_{1/2}$, h) were set to (0,0,0,1) and (1,2,8,3), respectively. The dynamic range was calculated with Eq. 18

$$d = \frac{S_m - S_0}{S_0} \quad (18)$$

Transformation of plasmids into *S. cerevisiae* BY4700

A single-colony–derived overnight seed culture of the host strain *S. cerevisiae* BY4700 was grown in YPD medium at 30°C and 220 rpm until it reached an optical density at 600 nm (OD_{600}) of 0.6. The cells were collected by centrifuging at 3000 rpm for 5 min at 4°C, washed with ice-cold sterile water, and resuspended in 100 mmol of lithium acetate buffer (100 mmol of lithium acetate per liter of Tris-EDTA buffer). Next, the cells were divided into aliquots in 1.5-ml sterile Eppendorf tubes, each containing 20 μl. Each tube was then supplemented with 80 μl of transformation mixture [58.6 μl of 50% (w/v) polyethylene glycol–4000, 7.7 μl of lithium acetate buffer (1 M), 9 μl of dimethyl sulfoxide, and 4.7 μl of preboiled salmon sperm single-stranded DNA (10 mg/ml)] and 1 μg of plasmids, mixed by vortexing and incubated at 30°C for 30 min before being transferred to a 42°C water bath for another 15 min. The cell transformants were then collected by centrifuging at 3000 rpm for 5 min. Each tube of cells was washed with 100 μl of 5 mM $CaCl_2$, centrifuged, resuspended with 100 μl of sterile water, and streaked onto SC-Ura plates.

Individual characterization of the dose-response relationships for malonyl–coenzyme A biosensors

Two variants (pCYC1-OP_TATA-UAS_FAC and pHSP12-TATA_OP-UAS_FAC) were obtained from library stock, and the other eight biosensor variants (pCYC1-OP_TATA-UAS_DDC, pCYC1-OP_TATA-UAS_EBC, pHSP12-TATA_OP-UAS_BDC, pCYC1-OP_TATA-UAS_BEC, pEXG1-N30_OP-UAS_FDA, pCYC1-TATA_OP-UAS_EAC, pCYC1-OP_TATA-UAS_FDC, and pCYC1-OP_TATA-UAS_BDC) were constructed via Golden Gate Assembly (table S3). After transformation of these plasmids into BY4700, the strains were inoculated into 48-well deep-well plates with 1 ml of SC-Ura medium (synthetic complete medium lacking uracil) in each well. After culturing for 12 hours at 30°C and 250 rpm, 2 μl of cerulenin solutions of six distinct concentrations (0, 0.5, 1.0, 1.5, 2.5, and 4 mg/ml) was added to the corresponding well. The strains were then cultured for another 12 hours. For sample preparation, cells were collected by centrifugation (4°C; 8000*g* for 10 min) and resuspended in prechilled phosphate-buffered saline (PBS) to an OD_{600} of 2. BY4700 was used as a negative control. BY4700/POT1-pTEF2-mCherry-tADH1 and BY4700/POT1-pCYC1-YPet-tPGK1 were used as positive controls for mCherry and YPet, respectively. These control samples were prepared the same way as above. The fluorescence intensities of the cells were characterized on an LSRFortessa (BD Biosciences). The double-positive area, named Q2, was determined by the control samples, as described in a previous work (*39*). For each sample, 100,000 events in the Q2 area were analyzed.

Construction of the two-reporter plasmid

The two-reporter plasmid pMPTPV_dual_fluorescence was derived from the common vector pACYCDuet-1 by replacing the chloramphenicol resistance gene with the kanamycin resistance gene, replacing the *lacI* expression cassette with *mcherry*, and inserting the *sfgfp* cassette into the opposite strand of *mcherry*. The *mcherry* gene is controlled by a constitutive promoter, pL_M1-37 (*73*). To facilitate library construction, *sfgfp* is controlled by a variable region containing two Bsm BI restriction sites.

Feasibility verification of the two-reporter plasmid

Ten promoters were randomly selected from the EcoCyc database. The sequence of each promoter was defined as the 60 nt preceding the transcriptional start site. In addition, a medium-strength RBS, BBa_J61106 (TCTAGAGAAAGATAGGAGACACTAGT), was chosen for all strains to ensure the survival of strains [note that the combination of a strong promoter and a strong RBS is lethal to *E. coli* (*19*)]. After transformation of the plasmids containing different promoters into *E. coli* K12 MG1655, the resulting strains were individually cultured in LB medium containing kanamycin (initial OD_{600} = 0.02), with three biological replicates for each promoter. During cultivation, the growth rate and the expression of fluorescent protein were monitored by sampling and testing every hour. The OD_{600} was measured by a microplate reader (Tecan Infinite 200Pro), and the fluorescence intensity was assayed via FCM (BD LSRFortessa). The 10 strains showed no apparent differences in growth (fig. S11A), and the median value of sfGFP/mCherry remained stable after culturing for 16 hours (fig. S11B). Hence, we chose 16 hours for cultivation in subsequent experiments.

Preparation of library cells

The two plasmid libraries (the promoter library and combination library) were both ordered from Genewiz. We transformed each library into *E. coli* K12 MG1655 via the BTX Harvard ECM 630 High Throughput Electroporation System using optimized parameter settings (2.1 kV, 1 kilohm, 25 μF, 100 ng of plasmids/100 μl of competent cells). The transformed cells were incubated in LB medium (four times the volume of the competent cells) for 1 hour at 37°C for recovery and then plated onto 37 Φ150 LB agar plates containing kanamycin with an EasySpiral Pro (Interscience). Generally, ~10^4 single colonies per plate can be harvested with this protocol, enabling ~100 times coverage of the designated library. All colonies on the plates were rinsed off using sterile LB medium supplemented with kanamycin, collected by centrifugation (4°C; 8000*g* for 10 min) and then resuspended and thoroughly mixed to an

OD_{600} of 10 using fresh sterile LB medium containing kanamycin. The cell suspension was stored at −80°C in glycerol (final OD_{600} = 5).

Characterization of transcriptional impact on growth

We further tested whether different promoters have varying influences on growth. To this end, the stored promoter library cells were cultured in LB medium containing kanamycin (initial OD_{600} = 0.02) at 37°C for 16 hours. Cell samples were collected before and after cultivation, followed by plasmid extraction. The promoter regions were amplified by PCR (KAPA HiFi PCR Kit) using the Lib_F and Lib_R primers (table S2).

In a 50-μl reaction, 5 ng of the plasmid library was added as the PCR template. The sequencing library was prepared according to the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). Specifically, 30 ng of each purified PCR product was used to prepare the sequencing library. The DNA fragments were treated with NEBNext End Prep for end repair, 5′ phosphorylation and dA-tailing. Then, the fragments were ligated to NEBNext Adaptors, followed by USER Enzyme excision. Subsequently, the products were purified using NEBNext Sample Purification Beads and amplified by PCR for 6 cycles using the P5 and P7 primers. The products were again purified using NEBNext Sample Purification Beads, validated with an Agilent 2100 Bioanalyzer (Agilent Technologies), and quantified with a Qubit 4 Fluorometer (Invitrogen). Subsequently, the libraries were delivered to Novogene for sequencing.

Two biological replicates were analyzed in parallel in this experiment, which generated three NGS raw datasets. After the production of clean data by demultiplexing and removing adaptor regions, pairs of paired-end data were merged by Fast Length Adjustment of SHort reads (FLASH) script (*74*), and those reads without detected pairs were removed. Python scripts generated in house were then used to search for the "GGATN86ATGC" 94-mer in the sequencing reads (and the reverse complementary sequence), and those carrying mutations within the upstream (GGAT) or downstream (ATGC) flanking regions (4 nt each) were removed. The read counts were then adjusted using Eq. 19, where n is the number of sequencing libraries, to normalize the different sequencing depths of each library

$$\text{Normalization factor}_i = \frac{\sum_{i=1}^{n} \text{Read count}_i}{n \times \text{Read count}_i} \tag{19}$$

The library showed negligible variation in growth (fig. S12), which ensured the feasibility of using the library in subsequent Sort-Seq experiments.

Sort-Seq experiments

For both the promoter library and the combination library, a frozen glycerol stock of library cells (*E. coli* MG1655) was inoculated into 100-ml flasks containing 20 ml of LB medium with kanamycin to an initial OD_{600} of 0.02. Library cells were grown for 16 hours at 37°C and 220 rpm. The grown cells were transferred to fresh LB medium to an initial OD_{600} of 0.02 and grown again under the same conditions as above. A third round of dilution and growth was carried out to improve the expression stability of the fluorescent proteins. After growth, 500 μl of culture medium was chilled on ice immediately, and the cells were collected by centrifugation (4°C; 8000*g* for 10 min). The cells were resuspended in 500 μl of prechilled PBS. Each cell suspension was diluted 150-fold in PBS to prepare samples appropriate for sorting. Three biological replicates were prepared for Sort-Seq experiments.

Sorting was performed on a FACSAria SORP (BD Biosciences). Gating based on FSC-Area and SSC-Area was carried out to exclude noncell particles. The population in this gated area is referred to as P1. The fluorescence background noise for the two relevant wavelengths was calibrated using the blank untransfected MG1655 strain. Note that the blank strain was completely negative for both sfGFP and mCherry expression. The resulting double-positive area in the region corresponding to the FITC-Area (sfGFP) and the PE–Texas Red–Area (mCherry) is referred to as Q2 after P1. The prepared library cells were analyzed by cytometry to determine the density distribution contour of the fluorescence in Q2. Subsequently, in the histogram of sfGFP/mCherry, 12 bins were set to evenly split the overall distribution of the population in Q2 (figs. S15 and S23 and tables S4 and S5), referred to as P2 to P13 after Q2, to ensure that the number of cells in each bin was equal and improve the sorting efficiency. For calibration, ~2×10^6 unsorted cells in gate Q2 were first collected for each sample. In the main sorting process, the three replicates were individually sorted into the 12 bins as described above. Each sample was successively sorted three times using four-way sorting. In each of these sorting runs, cells falling in nonadjacent bins were collected to eliminate the conflicting events between them. Thus, P2, P5, P8, and P11 were simultaneously sorted in one run, as were P3, P6, P9, and P12. During sorting, the cell flow rate was kept at ~8000 events/s, and ~5×10^5 cells were collected in each bin.

The sorted cells were collected in 36 (3 samples × 12 bins) 5-ml polystyrene round-bottom centrifuge tubes (BD Falcon), each of which contained 500 μl of PBS. The entire contents of each tube were then each transferred to 100-ml flasks containing 20 ml of LB medium with kanamycin and cultured at 37°C for 7 hours. These cells were then subjected to plasmid library extraction. Together with the cells from gate Q2 of each of the three samples mentioned above, we obtained 39 plasmid libraries in total.

The promoter and RBS regions of *sfgfp* in each library were amplified through PCR (KAPA HiFi PCR Kit), using 12 8-nt barcoded primers to identify different sorting bins (primers sorting_P2 to _P13; table S2). The barcodes were designed according to the following principles. (i) The Levenshtein distance between every two barcodes was ≥4, (ii) the guanine-cytosine (GC) content was 20 to 80%, and (iii) there were no more than four consecutive identical bases. In a 25-μl PCR, 5 ng of plasmid library was added as a template. PCR products from the 12 bins for each sample were mixed, thus obtaining three sorted PCR products (from sorted library cells in different sorting bins for three samples) and three unsorted PCR products (from unsorted library cells in the Q2 gate for three samples). The resulting PCR products were analyzed and purified by electrophoresis. The sequencing libraries were prepared as described above and were then delivered to Novogene for sequencing.

Sort-Seq data processing

According to NGS data, the read count $R_{i,k}$ for variant$_i$ in bin$_k$ can be observed. In addition, by analyzing the cytometry data, we can obtain the ratio of cells sorted into bin$_k$ against all cells, which is denoted by C_k and listed in table S5. Hence, assuming an unbiased NGS quantification process, the proportion of variant$_i$ in bin$_k$ is Q_{ik}

$= R_{i,k}/R_k$. Here, R_k is the total read count for the NGS library derived from bin_k. Therefore, the probability of sorting variant_i into bin_k should be $P_{ik} = Q_{ik}C_k/\Sigma_{k=1}^{K} Q_{ik}C_k$.

For the library of *tnaC* variants and the malonyl–coenzyme A (CoA) biosensors, data processing was performed as described above. However, for the promoter library and the combination library, sorting errors did exist. Specifically, certain cells were mistakenly sorted into bins where they did not belong (fig. S16, A, B, D, and F). We ascribed this error to random screening due to cell adhesion with an error rate, ϵ, and hence, we modified P_{ik} as shown in Eq. 20.

$$\boldsymbol{P}'_i = [P'_{i1}, P'_{i2}, \ldots, P'_{iK}]^T = ReLU(\boldsymbol{E}^{-1}\boldsymbol{P}_i) \quad (20)$$

where

$$\boldsymbol{E} = (1 - K\varepsilon)\boldsymbol{I}_K + \varepsilon\boldsymbol{J}_K \quad (21)$$

Here, $\boldsymbol{I}_K$ is an identity matrix of size K, and $\boldsymbol{J}_K$ is a matrix of ones of size K. Last, the binned distribution was obtained by $P''_{ik} = P'_{ik}/\Sigma_{k=1}^{K} P'_{ik}$. For both the promoter and combination libraries, ϵ was set to 0.05, which made the binned distribution more precise (fig. S16, C, E, and G). This modification was crucial to ensure accurate calculation of the mean and SD values. Without implementing this adjustment, the calculated mean and SD would have been prone to error (fig. S17). Subsequently, the strains with $P''_{i1} > 0.5$ or $P''_{iK} > 0.5$ were ruled out, as they were not effectively sorted, resulting in a high degree of error in the calculation of expression noise. For the *tnaC* variants and the malonyl-CoA biosensor libraries, we did not exclude any data as the calculation responses are not substantially affected by the number of effective sorting bins. Moreover, to ensure the quality of the results, we eliminated the data with low consistency among replicates. Specifically, if the CV of the calculated mean or SD among biological replicates was greater than 0.5, then the related data were removed from the dataset.

Evaluation of the expression patterns of essential and nonessential genes

The essential genes were identified on the basis of a comprehensive pooled CRISPRi screening dataset (*75*) (threshold: fitness ≤ −6), whereas other genes were regarded as nonessential. We calculated the transcriptional strength of each gene. Specifically, the genes belonging to the operons closest to the downstream side of a promoter were considered to be driven by this promoter, and the transcription strength of a gene was calculated as the summation of its promoter strengths, as illustrated in fig. S21.

Identification of RBS strengths

To identify the translational strengths of the 13 RBSs, we defined each promoter strength as the expression strength in the promoter library, and then, we divided each expression strength from the combination library by the corresponding promoter strength. The median of the calculated results for combinations with the same RBS is defined as the translational strength of that RBS. The resulting order of the RBS strengths was as follows: apFAB872(0.15) < apFAB914(0.23) < apFAB864(0.36) < apFAB865(0.67) < apFAB927(0.75) < apFAB827(0.95) < apFAB894(1.02) < apFAB909(1.04) < apFAB839(1.36) < apFAB833(1.38) < apFAB834(1.41) < apFAB820(1.57) < apFAB916(2.13) (fig. S27).

Calculation of the expression noise residuals

Each noise-mean relationship was fitted by the empirical formula $CV^2 = C_1 + C_2/\text{Mean}$ (Fig. 5, A and B), and the residual of each expression pattern was calculated and sorted. In addition, to ensure the reliability of the analysis, we only considered the sequences with appropriate mean expression levels (>0.15 for the promoter library; >0.1 for the combination library). The 20 candidates from the promoter library and 25 from the combination library with the largest residues were reconstructed.

Identification of potential RpoD-binding sites

The DPinteract database contains computational predictions of possible RpoD-binding sites with 15- to 19-nt spacing in the *E. coli* genome (*48*). We searched these sequences in the promoter library and counted the number of potential RpoD-binding sites of each promoter. To avoid redundant results, we only accounted for independent hexamer pairs. Specifically, if the −35 or −10 region of two RpoD-binding sites overlapped, then we only considered the RpoD-binding site with a higher *z* score; otherwise, both of them were retained (fig. S34).

Supplementary Materials

This PDF file includes:
Figs. S1 to S35
Tables S1 to S6
Legends for data S1 to S5

Other Supplementary Material for this manuscript includes the following:
Data S1 to S5

REFERENCES AND NOTES

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
2. G. W. Li, X. S. Xie, Central dogma at the single-molecule level in living cells. *Nature* **475**, 308–315 (2011).
3. N. Friedman, L. Cai, X. S. Xie, Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 168302 (2006).
4. V. Shahrezaei, P. S. Swain, Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17256–17261 (2008).
5. H. Salman, N. Brenner, C.-K. Tung, N. Elyahu, E. Stolovicki, L. Moore, A. Libchaber, E. Braun, Universal protein fluctuations in populations of microorganisms. *Phys. Rev. Lett.* **108**, 238105 (2012).
6. J. Beal, Biochemical complexity drives log-normal variation in genetic expression. *Eng. Biol.* **1**, 55–60 (2017).
7. M. L. Heltberg, S. Krishna, M. H. Jensen, On chaotic dynamics in transcription factors and the associated effects in differential gene regulation. *Nat. Commun.* **10**, 71 (2019).
8. A. Raj, A. van Oudenaarden, Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
9. Y. Toya, H. Shimizu, Flux controlling technology for central carbon metabolism for efficient microbial bio-production. *Curr. Opin. Biotechnol.* **64**, 169–174 (2020).
10. Y. Xiao, C. H. Bowen, D. Liu, F. Zhang, Exploiting nongenetic cell-to-cell variation for enhanced biosynthesis. *Nat. Chem. Biol.* **12**, 339–344 (2016).
11. T. Fojo, S. Bates, Strategies for reversing drug resistance. *Oncogene* **22**, 7512–7523 (2003).
12. N. A. Saunders, F. Simpson, E. W. Thompson, M. M. Hill, L. Endo-Munoz, G. Leggatt, R. F. Minchin, A. Guminski, Role of intratumoural heterogeneity in cancer drug resistance: Molecular and clinical perspectives. *EMBO Mol. Med.* **4**, 675–684 (2012).
13. N. Q. Balaban, J. Merrin, R. Chait, L. Kowalik, S. Leibler, Bacterial persistence as a phenotypic switch. *Science* **305**, 1622–1625 (2004).

14. M. Kærn, T. C. Elston, W. J. Blake, J. J. Collins, Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
15. J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, J. S. Weissman, Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
16. A. Deloupy, V. Sauveplane, J. Robert, S. Aymerich, M. Jules, L. Robert, Extrinsic noise prevents the independent tuning of gene expression noise and protein mean abundance in bacteria. *Sci. Adv.* **6**, eabc3478 (2020).
17. O. K. Silander, N. Nikolic, A. Zaslaver, A. Bren, I. Kikoin, U. Alon, M. Ackermann, Correction: A genome-wide analysis of promoter-mediated phenotypic noise in *Escherichia coli*. *PLOS Genet.* **8**, e1002443 (2012).
18. E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, E. Segal, Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
19. S. Kosuri, D. B. Goodman, G. Cambray, V. K. Mutalik, Y. Gao, A. P. Arkin, D. Endy, G. M. Church, Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14024–14029 (2013).
20. B. Townshend, A. B. Kennedy, J. S. Xiang, C. D. Smolke, High-throughput cellular RNA device engineering. *Nat. Methods* **12**, 989–994 (2015).
21. N. Peterman, E. Levine, Sort-seq under the hood: Implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
22. E. Sharon, D. Van Dijk, Y. Kalma, L. Keren, O. Manor, Z. Yakhini, E. Segal, Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res.* **24**, 1698–1706 (2014).
23. C. G. de Boer, E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, A. Regev, Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).
24. W. L. Noderer, R. J. Flockhart, A. Bhaduri, A. J. Diaz de Arce, J. Zhang, P. A. Khavari, C. L. Wang, Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10**, 748 (2014).
25. N. Peterman, A. Lavi-Itzkovitz, E. Levine, Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. *Nucleic Acids Res.* **42**, 12177–12188 (2014).
26. S. T. Rutherford, J. S. Valastyan, T. Taillefumier, N. S. Wingreen, B. L. Bassler, Comprehensive analysis reveals how single nucleotides contribute to noncoding RNA function in bacterial quorum sensing. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E6038–E6047 (2015).
27. J. S. Hawkins, M. R. Silvis, B. M. Koo, J. M. Peters, H. Osadnik, M. Jost, C. C. Hearne, J. S. Weissman, H. Todor, C. A. Gross, Mismatch-CRISPRi reveals the co-varying expression-fitness relationships of essential genes in *Escherichia coli* and *Bacillus subtilis*. *Cell Syst.* **11**, 523–535.e9 (2020).
28. T. Wang, X. Zheng, H. Ji, T. L. Wang, X.-H. Xing, C. Zhang, Dynamics of transcription–translation coordination tune bacterial indole signaling. *Nat. Chem. Biol.* **16**, 440–449 (2020).
29. A. Schmitz, F. Zhang, Massively parallel gene expression variation measurement of a synonymous codon library. *BMC Genomics* **22**, 149 (2021).
30. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
31. Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, X. S. Xie, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
32. T. S. Ferguson, *Bayesian Density Estimation by Mixtures of Normal Distributions* (Academic Press Inc., 1983).
33. R. G. Marks, P. V. Rao, An estimation procedure for data containing outliers with a one-directional shift in the mean. *J. Am. Stat. Assoc.* **74**, 614–620 (1979).
34. M. Aitkin, G. T. Wilson, Mixture models, outliers, and the EM algorithm. *Technometrics* **22**, 325–331 (1980).
35. R. J. Beckman, R. D. Cook, Outlier ... ... s. *Technometrics* **25**, 119–149 (1983).
36. A. A. K. Nielsen, B. S. Der, J. Shin, P. Vaidyanathan, V. Paralanov, E. A. Strychalski, D. Ross, D. Densmore, C. A. Voigt, Genetic circuit design automation. *Science* **352**, aac7341 (2016).
37. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks. arXiv.1406.2661 [stat.ML] (10 June 2014).
38. J. K. Rogers, N. D. Taylor, G. M. Church, Biosensor-based engineering of biosynthetic pathways. *Curr. Opin. Biotechnol.* **42**, 84–91 (2016).
39. Y. Zhou, Y. Yuan, Y. Wu, L. Li, A. Jameel, X. H. Xing, C. Zhang, Encoding genetic circuits with DNA barcodes paves the way for machine learning-assisted metabolite biosensor response curve profiling in yeast. *ACS Synth. Biol.* **11**, 977–989 (2022).
40. J. Ansel, H. Bottin, C. Rodriguez-Beltran, C. Damon, M. Nagarajan, S. Fehrmann, J. François, G. Yvert, Cell-to-cell stochastic variation in gene expression is a complex genetic trait. *PLOS Genet.* **4**, e1000049 (2008).
41. E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, A. Van Oudenaarden, Regulation of noise in the expression of a single gene. *Nat. Genet.* **31**, 69–73 (2002).
42. I. M. Keseler, S. Gama-Castro, A. Mackie, R. Billington, C. Bonavides-Martínez, R. Caspi, A. Kothari, M. Krummenacker, P. E. Midford, L. Muñiz-Rascado, W. K. Ong, S. Paley, A. Santos-Zavaleta, P. Subhraveti, V. H. Tierrafría, A. J. Wolfe, J. Collado-Vides, I. T. Paulsen, P. D. Karp, The EcoCyc database in 2021. *Front. Microbiol.* **12**, 711077 (2021).
43. D. Barrios, C. Prieto, D3GB: An interactive Genome Browser for R, Python, and WordPress. *J. Comput. Biol.* **24**, 447–449 (2017).
44. S. Proshkin, A. Rachid Rahmouni, A. Mironov, E. Nudler, Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**, 504–508 (2010).
45. G. E. Johnson, J. B. Lalanne, M. L. Peters, G. W. Li, Functionally uncoupled transcription-translation in *Bacillus subtilis*. *Nature* **585**, 124–128 (2020).
46. V. K. Mutalik, J. C. Guimaraes, G. Cambray, C. Lam, M. J. Christoffersen, Q. A. Mai, A. B. Tran, M. Paull, J. D. Keasling, A. P. Arkin, D. Endy, Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
47. S. Lisser, H. Margalit, Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res.* **21**, 1507–1516 (1993).
48. K. Robison, A. M. McGuire, G. M. Church, A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284**, 241–254 (1998).
49. H. Feng, Y. Yuan, Z. Yang, X.-H. Xing, C. Zhang, Genome-wide genotype-phenotype associations in microbes. *J. Biosci. Bioeng.* **132**, 1–8 (2021).
50. D. Yang, S. Y. Park, Y. S. Park, H. Eun, S. Y. Lee, Metabolic engineering of *Escherichia coli* for natural product biosynthesis. *Trends Biotechnol.* **38**, 745–765 (2020).
51. H.-J. Chang, P. L. Voyvodic, A. Zúñiga, J. Bonnet, Microbially derived biosensors for diagnosis, monitoring and epidemiology. *J. Microbial. Biotechnol.* **10**, 1031–1035 (2017).
52. B. L. Sabatini, L. Tian, Imaging neurotransmitter and neuromodulator dynamics in vivo with genetically encoded indicators. *Neuron* **108**, 17–32 (2020).
53. N. Muranaka, V. Sharma, Y. Nomura, Y. Yokobayashi, An efficient platform for genetic selection and screening of gene switches in *Escherichia coli*. *Nucleic Acids Res.* **37**, e39 (2009).
54. J. C. Liang, A. L. Chang, A. B. Kennedy, C. D. Smolke, A high-throughput, quantitative cell-based screen for efficient tailoring of RNA device activity. *Nucleic Acids Res.* **40**, e154 (2012).
55. A. Eldar, M. B. Elowitz, Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
56. K. S. Farquhar, D. A. Charlebois, M. Szenk, J. Cohen, D. Nevozhay, G. Balázsi, Role of network-mediated stochasticity in mammalian drug resistance. *Nat. Commun.* **10**, 2766 (2019).
57. S. Hooshangi, S. Thiberge, R. Weiss, Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 3581–3586 (2005).
58. J. L. Payne, A. Wagner, The causes of evolvability and their evolution. *Nat. Rev. Genet.* **20**, 24–38 (2019).
59. Z. Bódi, Z. Farkas, D. Nevozhay, D. Kalapis, V. Lázár, B. Csörgő, Á. Nyerges, B. Szamecz, G. Fekete, B. Papp, H. Araújo, J. L. Oliveira, G. Moura, M. A. S. Santos, T. Székely, G. Balázsi, C. Pál, Phenotypic heterogeneity promotes adaptive evolution. *PLoS Biol.* **15**, e2000644 (2017).
60. M. Schmutzer, A. Wagner, Gene expression noise can promote the fixation of beneficial mutations in fluctuating environments. *PLoS Comput. Biol.* **16**, e1007727 (2020).
61. A. Aranda-Díaz, K. Mace, I. Zuleta, P. Harrigan, H. El-Samad, Robust synthetic circuits for two-dimensional control of gene expression in yeast. *ACS Synth. Biol.* **6**, 545–554 (2017).
62. M. Mundt, A. Anders, S. M. Murray, V. Sourjik, A system for gene expression noise control in yeast. *ACS Synth. Biol.* **7**, 2618–2626 (2018).
63. D. Benzinger, M. Khammash, Pulsatile inputs achieve tunable attenuation of gene expression variability and graded multi-gene regulation. *Nat. Commun.* **9**, 3521 (2018).
64. K. P. Gerhardt, S. D. Rao, E. J. Olson, O. A. Igoshin, J. J. Tabor, Independent control of mean and noise by convolution of gene expression distributions. *Nat. Commun.* **12**, 6957 (2021).
65. V. Chauhan, M. N. M. Bahrudeen, C. S. D. Palma, I. S. C. Baptista, B. L. B. Almeida, S. Dash, V. Kandavalli, A. S. Ribeiro, Analytical kinetic model of native tandem promoters in *E. coli*. *PLoS Comput. Biol.* **18**, e1009824 (2022).
66. R. Liu, M. C. Bassalo, R. I. Zeitoun, R. T. Gill, Genome scale engineering techniques for metabolic engineering. *Metab. Eng.* **32**, 143–154 (2015).
67. G. Chalancon, C. N. J. Ravarani, S. Balaji, A. Martinez-Arias, L. Aravind, R. Jothi, M. M. Babu, Interplay between gene expression noise and regulatory network architecture. *Trends Genet.* **28**, 221–232 (2012).
68. C. J. Hartline, A. C. Schmitz, Y. Han, F. Zhang, Dynamic control in metabolic engineering: Theories, tools, and applications. *Metab. Eng.* **63**, 126–140 (2021).
69. N. M. Belliveau, S. L. Barnes, W. T. Ireland, D. L. Jones, M. J. Sweredoski, A. Moradian, S. Hess, J. B. Kinney, R. Phillips, Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4796–E4805 (2018).

70. T. A. Whitehead, A. Chevalier, Y. Song, C. Dreyfus, S. J. Fleishman, C. De Mattos, C. A. Myers, H. Kamisetty, P. Blair, I. A. Wilson, D. Baker, Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).
71. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).
72. A. Adams, D. E. Gottschling, C. A. Kaiser, T. Stearns, *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual, 1997 Edition* (Cold Spring Harbor Laboratory Press, 1998).
73. J. Lu, J. Tang, Y. Liu, X. Zhu, T. Zhang, X. Zhang, Combinatorial modulation of galP and glk gene expression for improved alternative glucose utilization. *Appl. Microbiol. Biotechnol.* **93**, 2455–2462 (2012).
74. T. Magoč, S. L. Salzberg, FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
75. T. Wang, C. Guan, J. Guo, B. Liu, Y. Wu, Z. Xie, C. Zhang, X.-H. Xing, Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nat. Commun.* **9**, 2475 (2018).

Acknowledgments: We would like to thank X. Zheng and Y. Zhou for sharing their experimental skills and data. We thank Y. Wu for construction of the pMPTPV plasmid. We thank B. Yu for help with the FACS experiments and F. Liu for help with NGS library construction. We thank F. Zhang (Washington University in St. Louis) for critical discussions regarding this work. **Funding:** This work was supported by the National Key Research and Development Program of China (2019YFA0904800), the National Natural Science Foundation of China (U2032210), the Foshan-Tsinghua Innovation Special Fund (THFS01), the Open Project Funding of State Key Laboratory of Biocatalysis, and the Enzyme Engineering of Hubei University (SKLBEE2020001)
Author contributions: Conceptualization: H.F. and C.Z. Methodology: H.F. Investigation: H.F. and F.L. Experiment: H.F. and F.L. Supervision: C.Z., A.-p.Z., and X.-h.X. Writing—original draft: H.F., F.L., and C.Z. Writing—review and editing: H.F., F.L., C.Z., T.W., and A.-p.Z. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Raw NGS data of Sort-Seq have been deposited into the NCBI Sequence Read Archive with BioProject accession number PRJNA800535. The source code, cytometry data, and plasmid maps related to this work can be accessed via Zenodo (https://doi.org/10.5281/zenodo.8139074) or GitHub repository (https://github.com/fenghuibao/dSort-Seq).

Submitted 4 January 2023
Accepted 6 October 2023
Published 8 November 2023
10.1126/sciadv.adg5296