ACS SyntheticBiology

pubs.acs.org/synthbio

Research Article

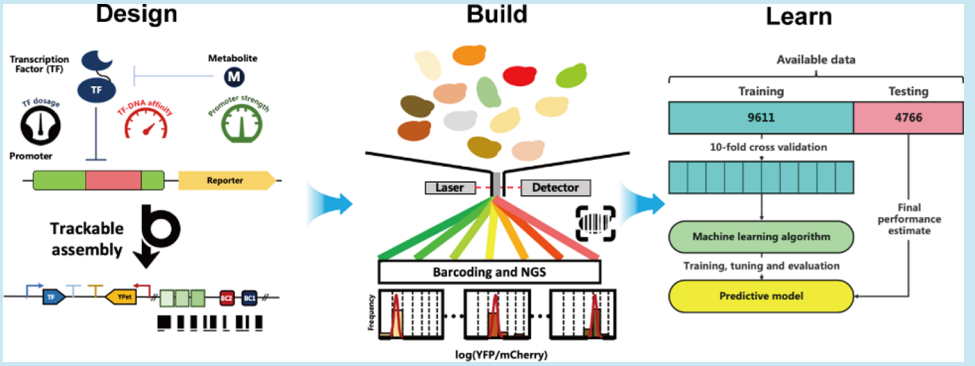
Encoding Genetic Circuits with DNA Barcodes Paves the Way for Machine Learning-Assisted Metabolite Biosensor Response Curve Profiling in Yeast

Yikang Zhou,[§] Yaomeng Yuan,[§] Yinan Wu, Lu Li, Aysha Jameel, Xin-Hui Xing, and Chong Zhang*

Cite This: *ACS Synth. Biol.* 2022, 11, 977−989

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Genetically encoded biosensors are valuable tools used in the precise engineering of metabolism. Although a large number of biosensors have been developed, the fine-tuning of their dose−response curves, which promotes the applications of biosensors in various scenarios, still remains challenging. To address this issue, we leverage a DNA trackable assembly method and fluorescence-activated cell sorting coupled with next-generation sequencing (FACS-seq) technology to set up a novel workflow for construction and comprehensive characterization of thousands of biosensors in a massively parallel manner. An FapR-*fapO*-based malonyl-CoA biosensor was used as proof of concept to construct a trackable combinatorial library, containing 5184 combinations with 6 levels of transcription factor dosage, 4 different operator positions, and 216 possible upstream enhancer sequence (UAS) designs. By applying the FACS-seq technique, the response curves of 2632 biosensors out of 5184 combinations were successfully characterized to provide large-scale genotype−phenotype association data of the designed biosensors. Finally, machine-learning algorithms were applied to predict the genotype−phenotype relationships of the uncharacterized combinations to generate a panoramic scanning map of the combinatorial space. With the assistance of our novel workflow, a malonyl-CoA biosensor with the largest dynamic response range was successfully obtained. Moreover, feature importance analysis revealed that the recognition sequence insertion scheme and the choice of UAS have a significant impact on the dynamic range. Taken together, our pipeline provides a platform for the design, tuning, and profiling of biosensor response curves and shows great potential to facilitate the rational design of genetic circuits.

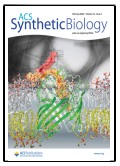
KEYWORDS: *genetically encoded biosensor, trackable assembly, machine learning, dose−response curve, yeast*

INTRODUCTION

Living organisms have a variety of mechanisms to sense and respond to environmental signals (i.e., chemical compounds, pH, temperature, etc.) by dynamically regulating their genetic expression networks. Harnessing this ability, genetically encoded biosensors were developed by mimicking natural genetic regulation networks through the assembly of basic biological parts like promoters, ribosome binding sites, operators, reporters, etc. in genetic circuits to recognize the analytes and transduce their signals into a measurable output. Given the advantages of portability, low cost, flexibility of design, and independence of special equipment, genetically encoded biosensors have gained extensive interest in many fields such as high-throughput screening,[1,2] dynamic control of microbial cell factories,[3,4] medical diagnosis, and health monitoring.[5]

The dose−response curve, which shapes how biosensors respond to specific signals and maps genetic circuits to the functions, is crucial for the utilization of genetically encoded

Received: November 27, 2021
Published: January 28, 2022



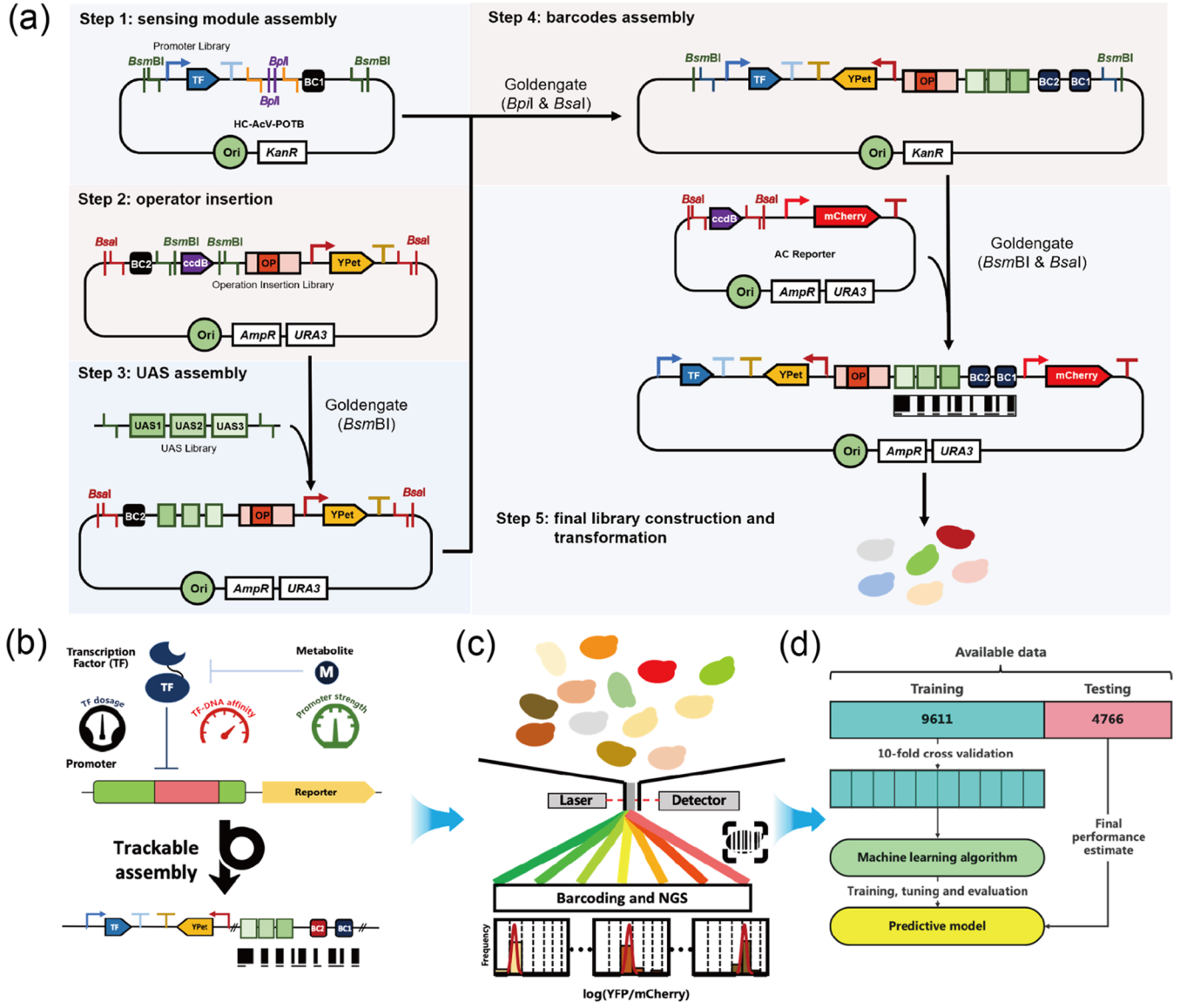


Figure 1. Massively parallel build-test-learning steps for the metabolite biosensor in yeast. (a) Overview of the encoding workflow for TF-based metabolite biosensors. The vertical lines represent the recognition sites of the enzyme and the elbow-shaped polylines are the enzyme digestion sites or sticky ends. The "T" logo is the terminator and the polyline arrow refers to the promoter. (b) Trackable metabolite biosensor library, which condensed high-order genotype information in barcodes, would make it possible for NGS to determine each design in the library. (c) Schematics of FACS-seq-guided high-throughput metabolite biosensor response profiling. After treating cells with different concentrations of inducers, the cells harboring the library were sorted into several bins according to their responses. The plasmid library in each bin was extracted, barcoded, and sequenced. The profile of read count numbers across all bins was used to dissect the μ value of each variant under the treated conditions. (d) Schematics of machine-learning progress. Available data would be split into a training set and a testing set. The training set and the test data would be used to fit and test the model, respectively.

biosensors under specific usage scenarios. For a transcription factor (TF)-based biosensor, the dose−response curve can be affected by multiple factors, such as the expression strength of the reporter (leaky expression, dynamic range, and sensitivity) and the concentration of TF (limit of detection). Since the influences of these factors are coupled with each other, fine-tuning of the dose−response curves always require a precise adjustment of each affecting factor, which is time-consuming and laborious. With the emergence of DNA assembly technologies,[6−11] the ability to construct the genetic circuits with varieties of constituting biological parts has increased exponentially, which has largely expanded the combinatorial space of the constructed biosensor library. However, due to the "combinatorial explosion" problem, full factorial searches of the dose−response curves of the combinatorial space are infeasible with a traditional trial-and-error manner.

To address this issue, we adopted a massively parallel profiling method, namely, fluorescence-activated cell sorting coupled with next-generation sequencing (FACS-seq),[12,13] to generate accurate dose−response curves for thousands of biosensors. However, due to the limitation of the NGS read length (<150 bp), it is hard to associate high-order genetic combinations with its corresponding phenotypes.[14] To solve this, here, we developed a novel biological part assembly workflow to encode genetic circuits with short DNA barcodes, which makes sure the one-to-one correspondence of barcodes and the combinations of biological parts. Since the sorting and sequencing processes cannot cover the entire combinatorial

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

space, the generated genotype−phenotype association maps were further used to train a machine-learning algorithm to fit predictive models. Taken together, here, we proposed a framework by combining high-throughput experimental data and deep learning algorithms, which provides a scalable platform for the fine-tuning of biosensors in a customized and low-cost manner.

Given that it is very difficult to develop TF-based biosensors in eukaryotes, as proof of concept, we applied the novel workflow for the fine-tuning of dose−response curves of malonyl-CoA biosensors based on the FapR-*fapO* system in *Saccharomyces cerevisiae* (*S. cerevisiae*). We constructed 5155/5184 biosensors using the trackable assembly workflow, including a malonyl-CoA biosensor with the largest dynamic response range that has been reported so far in *S. cerevisiae*. Overall, 2632 out of 5155/5184 biosensors were characterized via FACS-seq,[12,13] and 1123 were successfully predicted by the machine-learning algorithm. Key features affecting the dose−response characteristics were also deciphered.

RESULTS

Encoding Metabolite Biosensors with DNA Barcodes Enables FACS-seq to Track Combinatorial Information. We selected six promoters of different relative activities (*pGPD, pENO2, pHSP12, pEXG1, pCYC1, pULI1*) to construct a sensing module with different TF dosages. In the reporter module, the minimal yeast synthetic promoter with a significantly shortened sequence length, high robustness, high orthogonality, and relative activity[15] was used to drive the expression of YPet. However, the biosensor characterization at the single-cell level is interfered by gene expression noise, which arises from both extrinsic (i.e., cell size and shape, plasmid copy number, cell cycle stage) and intrinsic (i.e., fluctuation in numbers of DNA, RNA, enzymes, TFs) noise.[16,17] As suggested in previous studies, extrinsic noise is the predominant source of gene expression noise in yeast.[17] Therefore, the mCherry protein driven by a constitutive promoter was also integrated into the final plasmid library and its expression level was used as an internal standard to reduce the influence of extrinsic noise[12,18] (Figure S1). We then took the ratio of yellow and red fluorescence intensity as the corrected expression of YPet for further experiments. The minimal synthetic promoter contains an AT-rich spacer sequence, an upstream enhancer sequence (UAS), a core element involving the TATA box, the transcription start site (TSS), and a 30-nucleotide spacer sequence (N30).[15] To explore the effects of operator insertion positions in the core element of the synthetic minimal yeast promoter, four different insertion schemes were designed as follows: (1) directly downstream of the TATA box (TATA_OP), (2) 1 bp before the TATA box (OP_TATA), (3) 1 bp before and directly downstream of the TATA box (OP_TATA_OP), and (4) directly downstream of the N30 sequence (N30_OP). As UASs would significantly affect the reporter promoter strength, six 10 bp minimal UASs (UAS_A, UAS_B, UAS_C, UAS_D, UAS_E, UAS_F)[15] were linked in tandem to establish high-strength, short promoters in yeast, making $6 \times 6 \times 6 = 216$ possible UAS designs. Thus, the combined space contained $6 \times 4 \times 6 \times 6 \times 6 = 5184$ possible designs (Table S1). Each combination is named according to the rule "TF promoter−operator sequence insertion site-UAS sequence (AAA to FFF)", such as pGPD-TATA_OP-UAS_AAA.

The key of encoding is to assemble the genetic elements into a genetic circuit and their corresponding barcodes into a contiguous stretch of DNA at the same time (Figure 1). For the metabolite biosensor, we designed the Barcode1 (BC1) to record the information about the promoter of the TF and the Barcode2 (BC2) to record the information about the operator insertion position. The designed molecular barcode is a 7 bp DNA fragment. There may be base deletions, insertions, or mutations in the subsequent PCR and sequencing processes, which might cause confusion between molecular barcodes. Therefore, the Levenshtein distance between any two molecular barcodes was designed to be greater than or equal to 3, which means that the conversion between two molecular barcodes required at least three single-base passes deletion, insertion, or mutation. BC1 and BC2 were placed upstream of the UASs of the synthetic promoter. The integrated barcodes (including UAS sequences, BC1 and BC2) recording each genetic combination would be amplified from pooled cell populations in an unbiased fashion and quantified using high-throughput sequencing to identify the distribution of each genetic combination.

CombiGEM used two pairs of isocaudomers (*Bam*HI and *Bgl*II, *Eco*RI and *Mfe*I) to achieve the scalable pooled assembly of barcoded high-order combinatorial genetic libraries.[19] It is known that the length of gene elements in *S. cerevisiae* is much longer than that in bacterial ones. For example, the promoters of *S. cerevisiae* can stretch hundreds of base pairs, leading to a higher probability to find restriction sites in these elements. If these isocaudomers were used for the assembly of metabolite biosensors in yeast, many genetic elements could not be applied. Therefore, type IIs restriction endonuclease is a better choice, which can decrease the probability to find restriction sites due to the reduced number of applied enzymes. In addition, the PCR process would also introduce unnecessary mutations and molecular recombination. It is necessary to avoid the PCR products in the assembly process and use standardized genetic elements instead to ensure the high fidelity of the whole assembly process.

We developed an encoding workflow for TF-based metabolite biosensors in yeast in five steps (Methods and Figures S2−S6): (1) sensing module assembly, (2) reporter module assembly—operator insertion, (3) reporter module assembly—UAS assembly, (4) metabolite biosensor assembly—barcodes assembly, and (5) transformation of the metabolite biosensors into yeast cells. The most important idea of the encoding workflow is that there remains a one-to-one correspondence between genetic circuits and barcodes. To verify this idea, single colonies were selected from the final yeast library to sequence the promoter that can drive the expression of TFs and the reporter promoter regions along with barcodes. The information of the barcodes and genetic elements was compared to determine whether they were completely matched or not. During the establishment of this encoding workflow, we first tried to mix all fragments to construct the mixed library. Overall, 6 of 128 colonies tested in the mixed library remained mismatched between the promoter that drives the expression of TFs and BC1 (Table S2). The barcode mismatch indicated that subsequent NGS genotyping analysis of the barcode region would not be able to track the combination information, which might seriously affect the reliability of subsequent data analysis results. Since all of the mismatches were found between the TF promoter and BC1, step 4 was further modified to avoid the barcode mismatch,

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

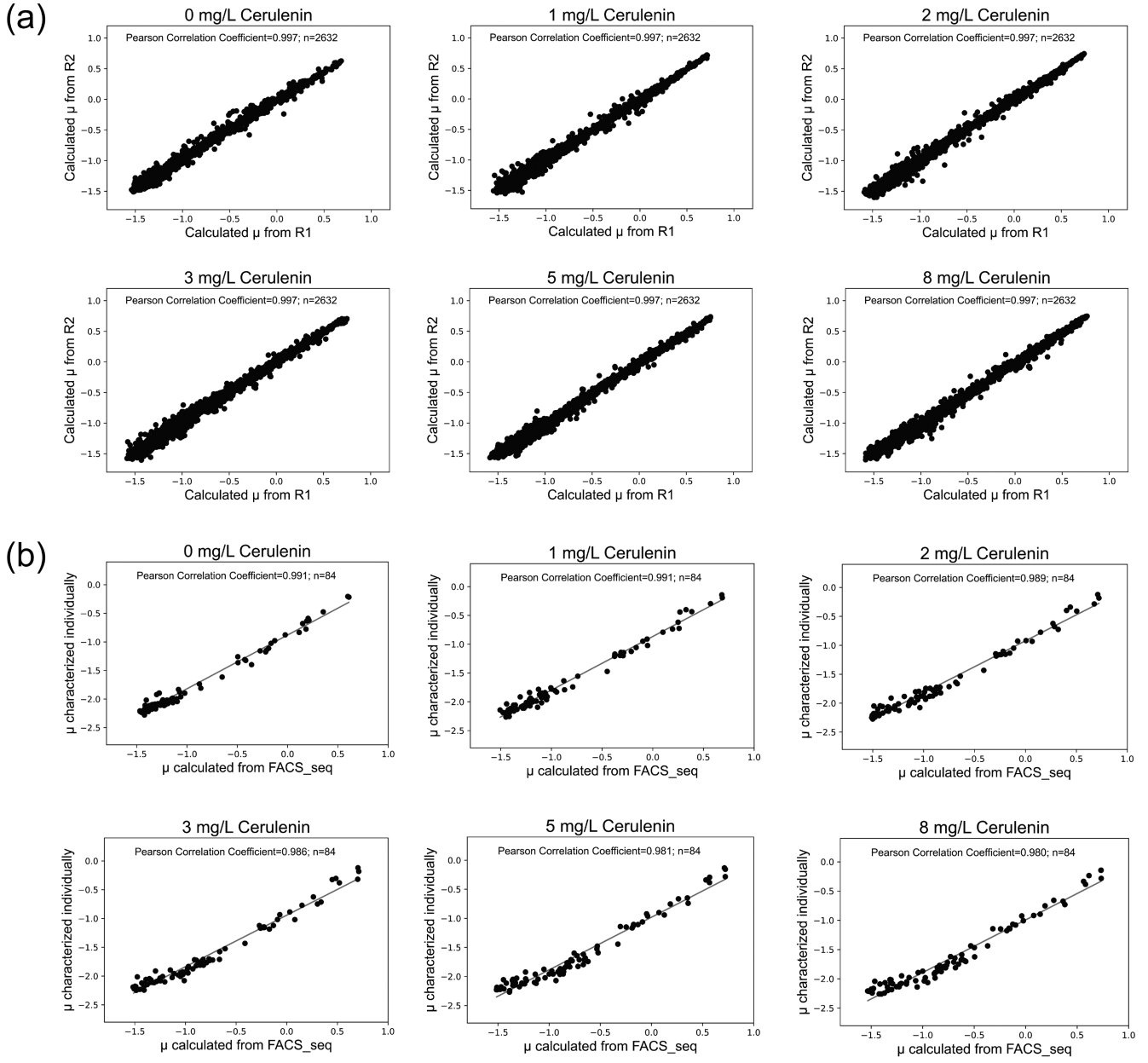


Figure 2. Results of the massively parallel characterization of the FapR-*fapO* based malonyl-CoA biosensors. (a) FACS-seq revealed strong consistency between biological replicates in the malonyl-CoA biosensor response under all experimental conditions (all six: n = 2632; Pearson correlation coefficient (PCC) = 0.997). (b) Individually analyzed responses of 84 designs by cytometry were highly correlated with those inferred from the FACS-seq approach (all six: n = 84; PCC > 0.980). Each point represents an individual design in (a, b).

where the six transcription units constructed in step 1 were separated to build a library. Not a single colony remained mismatched among the tested 133 colonies from the separated library (Table S3). The UAS mutations that existed in both the mixed library (4/128) and the separated library (8/133) could be detected in the NGS results, so there would be no interference with subsequent data analysis.

Massively Parallel Characterization of the Encoded Metabolite Biosensor Library by FACS-seq. We constructed the FapR-*fapO* regulatory system-based malonyl-CoA biosensor via the encoding workflow. Since it is difficult to change the intracellular malonyl-CoA concentration by malonyl-CoA feeding, different levels of cerulenin were instead added to positively regulate it.[20] We established a reliable FACS-seq based testing process using the encoded metabolite biosensors (Figure 1c). The encoded malonyl-CoA biosensor library was sorted into eight channels (termed as P2∼P9) according to the fluorescence intensity, and the abundance of each combination in each channel was determined by NGS. Out of 5184 designs, 5155 were successfully detected via NGS analysis. Using the FACS and NGS data, the μ values ($\log_{10}$(YPet/mCherry)) of each combination were calculated to realize the massively parallel genotype−phenotype association. We characterized the encoded library at six different concentrations of cerulenin (0, 1, 2, 3, 5, 8 mg/L) and left out the combinations whose total read counts in P2−P9 were less than 250. As a result, 16 753 pieces of data were successfully obtained in total, and 2632 designs were effectively estimated in all six concentrations (accounting for 50.8% of the theoretical library size). In all six experiments with different

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

Table 1. Prediction Effects of Different Machine-Learning Models

	training set			testing set		
model	MAE	MSE	PCC	MAE	MSE	PCC
linear regression	0.141	0.03375	0.95572	0.144	0.03428	0.95800
ridge regression	0.141	0.03375	0.95571	0.144	0.03428	0.95800
Lasso regression	0.141	0.03375	0.95571	0.144	0.03427	0.95801
support vector regression	0.059	0.00680	0.99138	0.057	0.00550	0.99348
random forest regression	0.013	0.00037	0.99953	0.062	0.00963	0.98839
gradient boosting regression	0.053	0.00599	0.99230	0.058	0.00647	0.99221
XGBoost regression	0.005	0.00005	0.99994	0.047	0.00586	0.99300

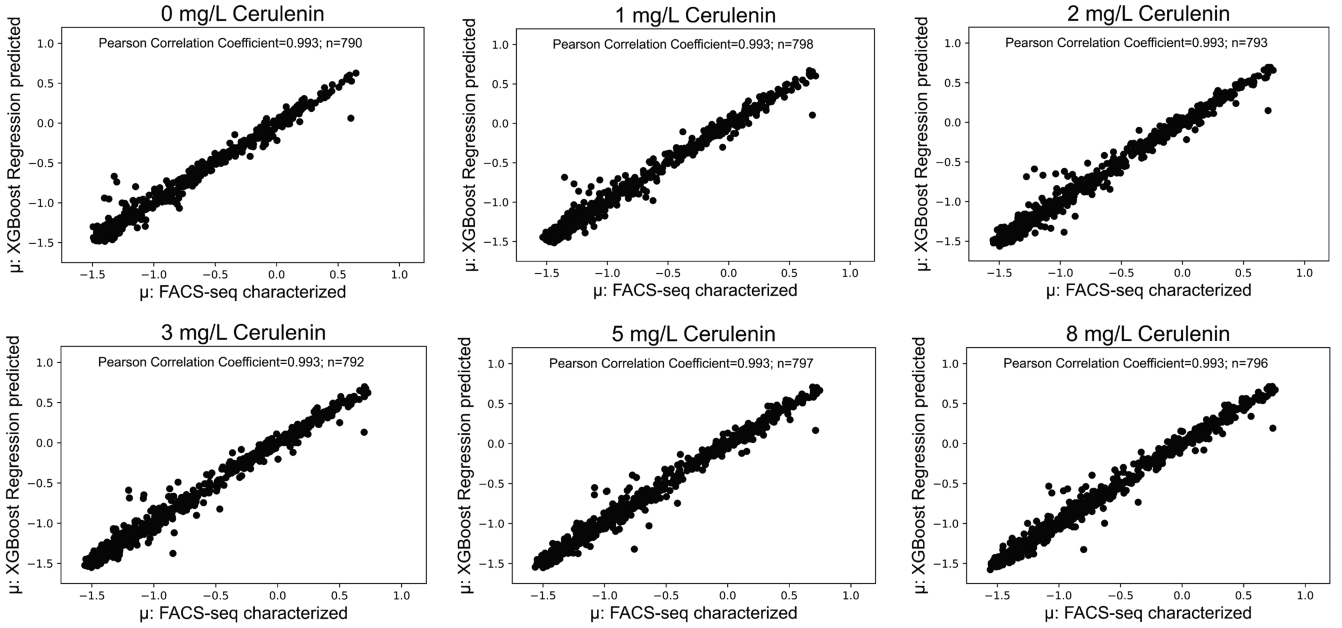


Figure 3. Prediction performance of XGBoost regression on the testing set (all six: PCC = 0.993). The horizontal axis and the vertical axis showed the μ values obtained by FACS-seq characterization and XGBoost regression, respectively.

concentrations of cerulenin, the two biological replicates were performed and agreed well with each other regarding the μ values (Figure 2a). Furthermore, 84 combinations covering a wide range of response levels were individually characterized via cytometry to experimentally validate the empirically determined results. A high correlation (Pearson correlation coefficient > 0.98) was observed between the results of the two methods (Figure 2b), which underscored the reliability of the FACS-seq approach.

The proportion of combinations successfully characterized by FACS-seq in *pGPD* and *pENO2* separated libraries were 0.093 and 0.253, respectively, which were much lower than the other four separated libraries (0.652-0.705) (Table S4). The Gini coefficients of the distribution in eight channels for *pGPD* and *pENO2* libraries with no cerulenin addition were 0.990 and 0.961, much higher than the other four separated libraries (Table S5). It showed that the genotype distribution in these two groups was extremely uneven. In fact, among the 84 combinations individually characterized, there were only 3 *pENO2* combinations (pENO2-TATA_OP-UAS_ECC, pENO2-OP_TATA-UAS_BEC, pENO2-OP_TATA-UAS_-BEF) and 1 *pGPD* combination (pGPD-OP_TATA-UAS_FBC). The extremely uneven distribution of these two groups might be caused by the fact that the maintenance and expression of some biosensors added a heavy metabolic burden to the host strains and then affected cell growth. During the culture process, the combinations having less impact on cell growth reached a significant enrichment level, while the abundance of others was substantially reduced.

Machine Leaning-Assisted Panoramic Scanning of the Combinatorial Space-Obtained Biosensors with High Performance. FACS-seq enabled a massively parallel characterization of the biosensor library and the acquirement of high-quality genotype−phenotype association data. However, there were still a considerable number of combinations not evaluated effectively due to the limited throughput of FACS sorting and the low abundance of some biosensors in the combinatorial space. To realize a panoramic scanning of the combinatorial space, machine learning was used to predict the μ values of the biosensors that were not successfully characterized by FACS-seq.

Since the proportion of combinations successfully characterized by FACS-seq in *pGPD* and *pENO2* separated libraries was significantly lower than the other four separated libraries, we temporarily eliminated these combinations (299 combinations, 2376 pieces of data) in the subsequent machine-learning step. The remaining genotype−phenotype associated data (14377 pieces of data in total) were split into a training set (2/3, 9611 pieces of data) and a test set (1/3, 4766 pieces of data) randomly (Figure 1d).

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

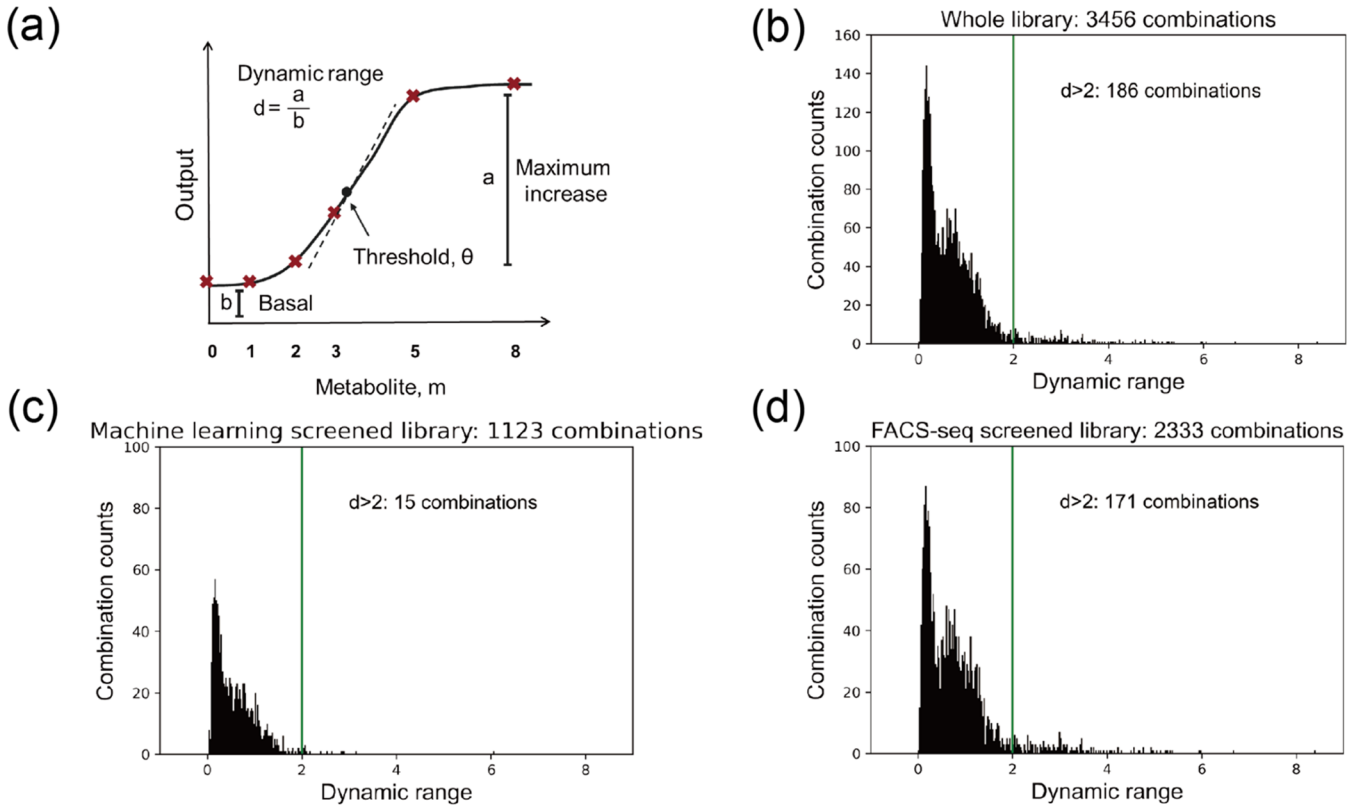


Figure 4. Distribution of the dynamic range. (a) Dose−response curve of metabolite biosensors and parameters: basal output (b); response threshold (θ), defined as the amount of the metabolite required for the 50% output expression relative to the baseline; maximal increase (a); dynamic range (d), defined as a/b; (b) distribution histogram of the dynamic response range of 2333 combinations, which were successfully characterized by FACS-seq in the four separate libraries; (c) distribution histogram of the dynamic response range of 1123 combinations, which were not characterized by FACS-seq but predicted by machine learning in the four separate libraries; and (d) distribution histogram of the dynamic response range of 3456 combinations, which were panoramically scanned in the four separate libraries.

First, we tried the linear regression model, and the mean absolute error (MAE) and the mean square error (MSE) between the model prediction result and the FACS-seq characterization result were large. However, the Pearson correlation coefficient (PCC) reached more than 0.95, suggesting a good correlation between them. Then, we used ridge regression and Lasso regression, in which regularization elements were introduced to optimize fitting, but the performance was not improved (Table 1). Support vector regression was then applied, and the PCC on the training set and the test set was above 0.99. Moreover, the MAE and MSE were significantly declined (Table 1) compared to the linear regression model.

Subsequently, we applied three commonly used machine-learning ensemble models—random forest regression, gradient boosting regression, and XGBoost regression—to further improve the prediction performance. After 10-fold cross-validation and the adjustment of model hyperparameters, these three ensemble models showed improved prediction performances. In particular, XGBoost regression performed best compared to all of the other models on the training set and the test set (Table 1). Although the prediction results on the test set showed a significantly higher MSE than the training set, still XGBoost regression maintained a high prediction performance, indicating the good generalization ability (Figure 3).

In the four separated libraries, FACS-seq obtained the μ values of 2333 combinations at six different concentrations of cerulenin. Since the errors of XGBoost regression on the training set were extremely small (Table 1), the μ values obtained by the machine-learning model and FACS-seq coincide with each other. XGBoost regression was applied for the remaining 1123 combinations that failed to obtain the μ values by FACS-seq at six different concentrations of cerulenin. Panoramic scanning was realized by performing the dose−response curve fitting on 3456 combinations in the four separated libraries (Figure 4). Figure 4b,c shows the distribution of 2333 combinations' and 1123 combinations' dynamic range. Both figures indicated that the dynamic ranges of most combinations in the library were small. In the four separated libraries, only 182 combinations (accounting for 5.3%) showed a dynamic response range larger than 2 (Figure 4d).

We reconstructed and characterized the top five combinations (DRTop5, Figure 5) with the largest predictive dynamic range, among which the FACS-seq failed to characterize pCYC1-TATA_OP-UAS_FFC due to its low abundance. According to the results of individual characterization, the dynamic ranges of DRTop5 were 3.57−6.38-fold. The response curves of DRTop5 obtained by prediction and individual characterization showed a good consistency (Figure 5).

Exploring the Capacity of TATA_OP-UAS_FAC Combinations. We noticed from the genotype−phenotype association data of DRTop5 that the dose−response curves of TATA_OP-UAS_FAC have the potential for a large

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

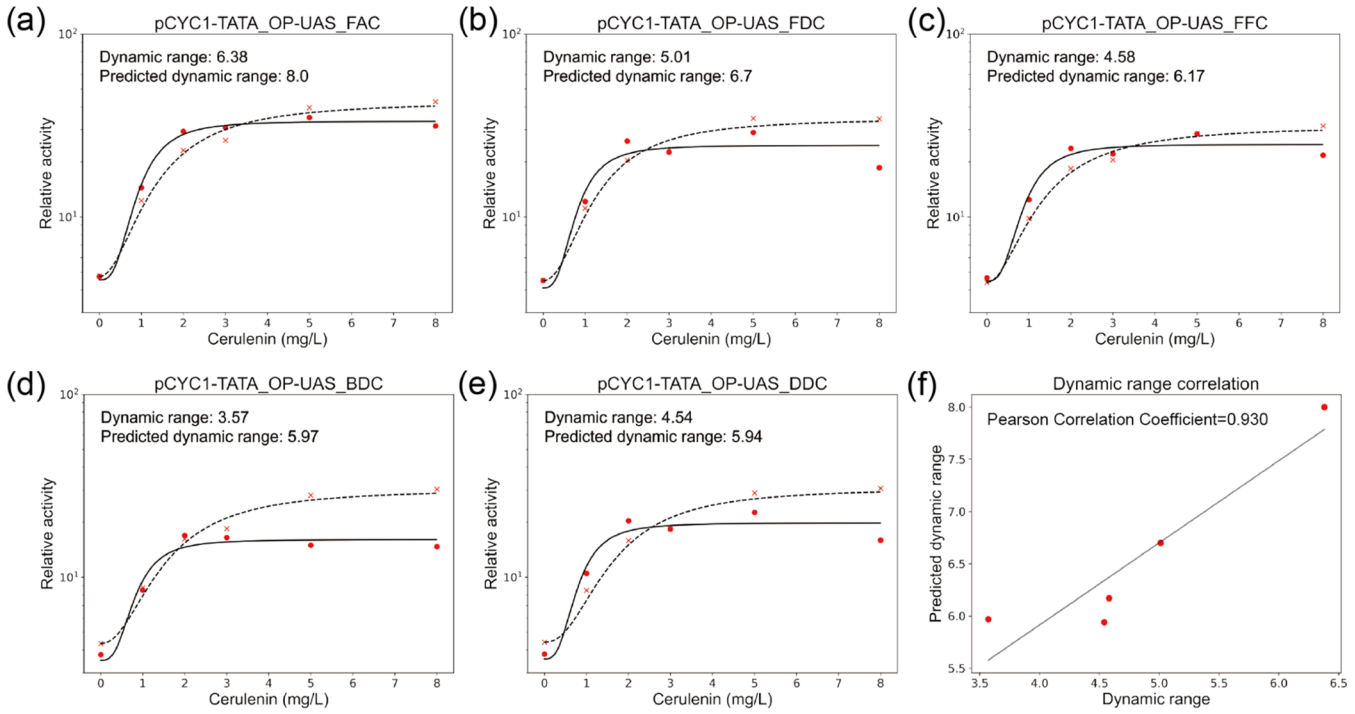


Figure 5. Predicted and individually characterized dynamic ranges of DRTop5. (a−e) Dose−responsive curves of DRTop5. The red dots and crosses are the μ values in eight concentrations obtained by individual characterization and XGBoost regression prediction, respectively. The solid lines and dashed lines are their corresponding response curves. (f) Predicted and individually characterized dynamic ranges show a high consistency (PCC = 0.93).

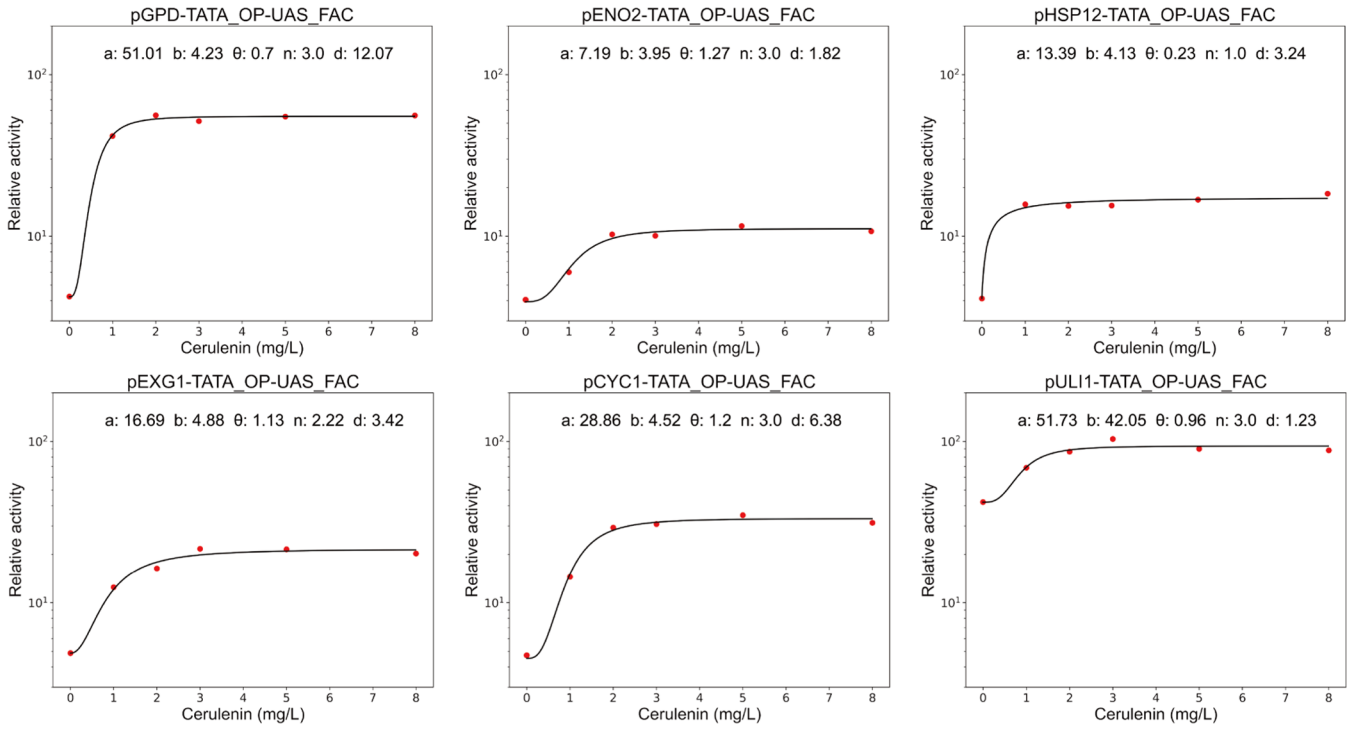


Figure 6. Response curves of the combinations of six promoters and TATA_OP-UAS_FAC. *b*: basal output, *a*: maximum increase value, θ: response threshold, *n*: Hill coefficient, and *d*: dynamic response range.

dynamic range. Therefore, the response curves of the combination of six promoters and TATA_OP-UAS_FAC were constructed and characterized individually. The dynamic range of pGPD-TATA_OP-UAS_FAC reached 12.07-fold (Figure 6). In *S. cerevisiae*, the reported largest dynamic range of the malonyl-CoA biosensor based on the FapR-*fapO* system was up to sevenfold.[21] Thus, we obtained the malonyl-CoA biosensor with the largest dynamic range in yeast cells so far. Compared with the other five combinations, the basal output of pULI1-TATA_OP-UAS_FAC was significantly

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

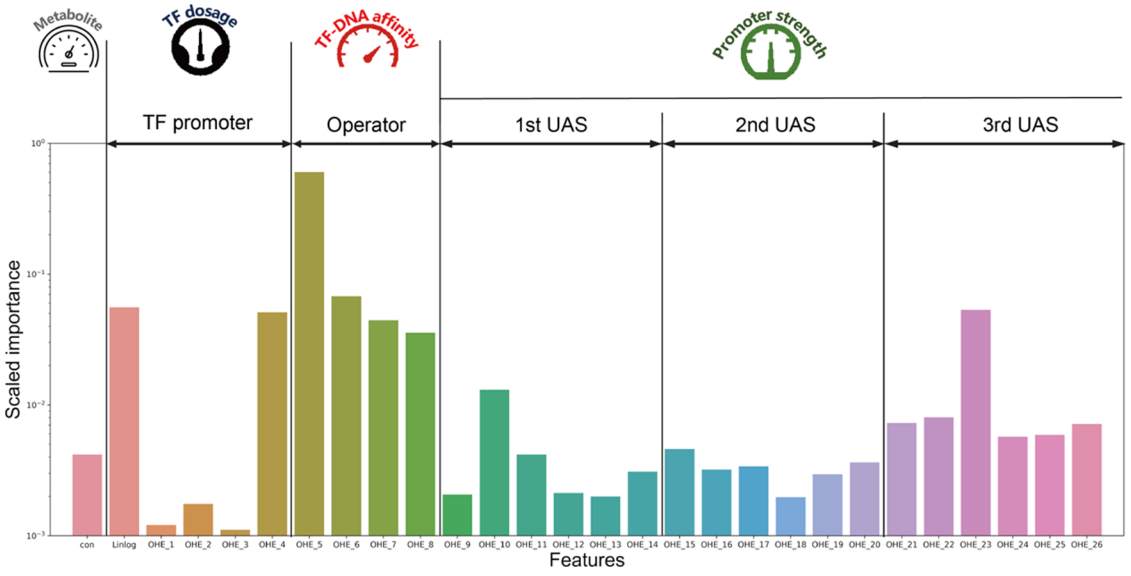


Figure 7. Feature importance of XGBoost regression. con: cerulenin concentration; linlog: linlog-transformed activity of the promoters (*pCYC1, pEXG1, pHSP12, pULI1*); OHE_1-OHE_4: one-hot encoded (OHE) features for the *fapR* gene expression driven by *pCYC1, pEXG1, pHSP12*, and *pULI1*; OHE_5-OHE_8: features for operator insertion schemes (N30_OP, OP_TATA, OP_TATA_OP, TATA_OP); OHE_9-OHE_14, OHE_15-OHE_20, and OHE_21-OHE_26: features for three linked (first, second, third) UAS sequences (UAS_A, UAS_B, UAS_C, UAS_D, UAS_E, UAS_F).

higher than that of other groups. Although its dynamic range was lower, the maximum increase value was equivalent to pGPD-TATA_OP-UAS_FAC, indicating a strong fluorescence output at high cerulenin concentrations.

Analysis of Features Using XGBoost Regression. XGBoost regression was applied for the FACS-seq characterized combinations and the importance of 28 key features was analyzed (Figure 7). First, the importance of the operator insertion schemes (OHE_5-OHE_8) was much higher than others, especially the OHE_5, which corresponds to the N30_OP insertion scheme, largely determined the μ values. However, when no cerulenin was added, the μ values of the N30_OP insertion scheme were significantly higher than the other three features, suggesting that this insertion scheme may result in a higher basal output of the biosensors (Figure S7). Second, the distribution of the third UAS showed high importance (OHE_21-OHE_26). When the third UAS was UAS_C, significantly higher output signals (μ values) were predicted in some reporter promoters. However, the same pattern was not observed in the second UAS (Figure S8). A certain strength of the reporter promoter was needed to ensure a detectable output signal. Therefore, when the yeast minimal synthetic promoters were utilized for the biosensor design, UAS_C was suggested to be used as the third UAS. Nevertheless, one should be aware that the abovementioned pattern might be related to the sequence of the TF recognition sequence itself and the insertion scheme. Third, the importance of the linlog activity of TF promoters was also higher than others (Figure 7). The YeastFab's promoter activity data characterized by fluorescent protein were used in the linlog activity feature, while TF promoter features (OHE_1-OHE_4) represented the intracellular FapR expression level. Therefore, the linlog activity feature was necessary although it was overlapped with the TF promoter features. Moreover, XGBoost regression predicted a significantly different μ value distribution of the FapR expression driven by *pULI1*, probably due to very low activity of *pULI1* (Figure S9). Finally, the feature of the cerulenin concentration showed low importance, which might be related to the poor response performance of most combinations in the combinatorial space. As the concentration of cerulenin increased, a right-shifted μ value distribution was predicted by machine learning (Figure S10), which was consistent with the results obtained by FACS-seq.

DISCUSSION

Large genetic engineering required the genetic circuits to coordinate the response, such as the turn on/off of metabolic valves at different growth stages. Although promising achievements have been made in the design and assembly of genetic circuits that perform correctly at a digital level, further optimization procedures are still required to obtain a desired function in the genetic circuits owing to the lack of ability to rationally predict the quantitative response in host cells. Since the biological parts that affect the response curves are coupled with each other, the optimization of the genetic circuits needs to adjust each part precisely, which is a time-consuming and laborious design-build-test cycle. The DNA assembly technologies have expanded the combinatorial space of the constructed genetic circuit library. However, it is difficult to panoramically scan the whole landscape of the response curves using the current characterization approaches. In this work, FACS-seq was combined with the DNA trackable assembly method to develop a novel workflow for the construction and characterization of numerous genetic circuits simultaneously, which provided a wider band for the design-build-test process. With a large amount of genotype−phenotype association data, machine-learning algorithms were further applied to predict the uncharacterized dose−response curves. Compared with the traditional trial-and-error process, our workflow provided a time-saving and low-cost alternative for the fine-tuning of genetic circuits.

The TF-based metabolite biosensors, which mainly exist in the prokaryotic kingdom, have emerged as an efficient tool for the development of microbial cell factories. Recently, methods

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

have been established to transplant them into the eukaryotic chassis, like yeast.[22,23] To further promote the application of TF-based biosensors in yeast, we chose the malonyl-CoA biosensor based on the FapR-*fapO* system in *S. cerevisiae* as the model to test our workflow. Malonyl-CoA is a key node in the metabolic pathway and an important precursor of 3-hydroxypropionic acid, flavonoids, fatty acids, and other compounds. The expression level of FapR, the insertion site of *fapO*, and the promoter strength of the output reporter gene can significantly affect the components of the biosensors. Therefore, their corresponding genetic elements were assembled and combinatorically optimized. To overcome the limitation of the length of NGS, the genetic circuits were encoded with DNA barcodes, which made sure the one-to-one correspondence of combinations of the barcodes and biological parts. With our five-step encoding workflow, a trackable combinatorial library containing 5155/5184 combinations was constructed. The FACS-seq technique successfully characterized the response curves of 2632 biosensors out of 5184 combinations and provided a large-scale genotype−phenotype association data of the designed biosensors. Machine-learning algorithms were then developed to predict uncharacterized dose−response curves and identify key features in the whole combinatorial library to generate a panoramic scanning map of the combinatorial space. With the assistance of our novel workflow, 3755 (2632 characterized + 1123 predicted) dose−response curves were obtained at a cost of $1.37 per curve (Table S6), and a malonyl-CoA biosensor with the largest dynamic response range was successfully acquired. Moreover, the feature importance analysis revealed that the recognition sequence insertion scheme and the choice of UAS have a great impact on the dynamic range. These results highlighted that our workflow has a great potential to speed up the tuning and application of the TF-based biosensors in yeast.

Compared with electronic engineering, the information foundation of a biological system design is still in the inception phase. In addition, the technologies for the utilization, manufacturing, and assembly of biological systems are still under development. The realization of "bottom-up" construction of synthetic biological systems relies on the development of a standard computing framework that is able to cover each stage in the DBTL cycle and also the storage and mining of information about the success or failure of synthetic biological systems. We explored the potential of machine learning to accelerate the DBTL cycle, showing that large-scale, high-quality test bandwidth could greatly improve the prediction efficiency of machine learning, which could also deepen the understanding of genetic circuit optimization.

Since our framework used a more versatile type IIs restriction enzyme and has a great ability to achieve a more efficient standardized assembly of genetic circuits, our work can serve as a platform for tuning and profiling the dose−response curves of more metabolites in a wide range of host cells. More broadly, we envision that more types of genetic circuits, e.g., AND, NOT, and OR logic, could be constructed using this encoding workflow (Figure S11).

Overall, our pipeline provides an efficient, affordable, and universal platform that enables the high-throughput scanning of dose−response curves and facilitates the rational design of the genetic circuits. Although the increased bandwidth of the DBT process paves the way for the machine learning-assisted profiling of genetic circuits, there is still a long way to go. Data utilization should be improved to decipher the characteristics of the sequences of biological parts in the future by extending this data-driven strategy from the feature extraction of combination information to the sequence features. In addition, more algorithms should be developed to support the prediction and customized design of genetic circuits. Ultimately, further optimization of the NGS library preparation in our workflow will probably lower its cost. Taken together, we are positive that the advancement of data science and synthetic biology assisted with our designed workflow will speed up the customized design of genetic circuits.

METHODS

DNA Manipulations. Plasmid extraction from *Escherichia coli* (*E. coli*) and DNA fragment purification were performed using kits from Omega Bio-Tek. The Solarbio Yeast Plasmid Extraction Kit was used to extract plasmids from yeast. PCR reactions were carried out using Q5 High-Fidelity DNA polymerase (NEB). Restriction enzyme FastDigest *Esp*3I (namely *Bsm*BI), *Eco*31I (namely *Bsa*I), *Bpi*I, and T4 DNA ligase were purchased from Thermo Scientific. All oligonucleotides were ordered from Genewiz. Molecular cloning was performed with *E. coli* DH10B (BioMed) as the host unless specially mentioned. For the plasmids containing the *ccdB* gene, molecular cloning was performed with *E. coli* DB3.1. To make annealed linkers, 4 μL of the complementary oligonucleotides (100 μM) was mixed in a 0.20 mL PCR tube with 4 μL of 5× annealing buffer (Beyotime, Haimen, Jiangsu, China) and 20 μL total volume of distilled water. The PCR tube was placed in a PCR block with the program: (1) 95 °C, 2:00; (2) 95 °C, 0.08, −0.1 °C per cycle; (3) GOTO step 2, 700 cycles; and (4) 4 °C ∞. Sanger sequencing provided by Genewiz was used to check the barcodes corresponding to the genotype information. The Phire Plant Direct PCR Master Mix was used to amplify the sequence from the yeast colonies directly. All strains, plasmids, and primers used in this work are summarized in Tables S7 and S8.

Cell Growth Conditions and Strain Construction. In all experiments, *E. coli* and yeast were grown at 37 and 30 °C, respectively. The *E. coli* strains were cultured in lysogeny broth (OXOID) containing 50 mg/L ampicillin (Ameresco) or 75 mg/L kanamycin (Ameresco) for plasmid maintenance. Yeast strains were cultured at 30 °C in a synthetic complete medium (SC) lacking uracil (SC-Ura) for selection. Cerulenin (C4643: 5 mg) was ordered from ABExBIO and dissolved in 1 mL of ethanol. The 5 mg/mL cerulenin stock solution was diluted to different concentrations (0.5, 1.0, 1.5, 2.5, 4 mg/mL) with ethanol.

Workflow of Encoding. *Step 1: Sensing Module Assembly.* YeastFab assembly synthesized the promoters (PRO), open-reading frames (ORF), and terminators (TER) as standardized biological parts using two different type IIs enzymes, namely, *Bsa*I and *Bsm*BI, which can be used to assemble the transcription units PRO-ORF-TER into a POT vector in a single-tube reaction.[24] We modified the sticky ends of the standardized terminator (Table S9) and constructed a New_HCKan_T_tCYC1 plasmid so that we could add BC1 after the PRO-ORF-TER. The annealed linker, which contained the BC1 and two *Bpi*I restriction sites, was added in the assembly reaction (Figure S2a) and PRO-ORF-TER-BC1 was then constructed due to the compatible sticky ends (Table S9). The *Bpi*I restriction sites introduced by the linker produced the compatible sticky ends for the insertion of the reporter module.

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

The standardized biological parts in YeastFab should avoid *Bsa*I and *Bsm*BI restriction sites. In our encoding strategy, these parts only need to avoid *Bpi*I restriction sites, showing that more standardized parts can be used. To achieve different TF dosages in cells, six promoters with different relative activities, *pGPD, pENO2, pHSP12, pEXG1, pCYC1*, and *pULI1*, were chosen to drive the expression of the gene *fapR*. Among them, the strongest promoter *pGPD* obtained 150 times higher relative activity than the weakest promoter *pULI1* (Table S10). These selected six promoters and their corresponding annealed linkers were added into the reaction system in Table S11 to construct POTB. The GoldenGate reaction program is shown in Table S12. As the recognition site of *Bpi*I exists on the POTB plasmid, transcription unit PRO-ORF-TER-BC1 was then transferred to the designed vector HC-AcV plasmid to construct HC-AcV-POTB (Figure S2b and Table S13).

Step 2: Reporter Module Assembly—Operator Insertion. In the reporter module, synthetic minimal yeast promoters were used to drive the expression of YPet. To explore the effects of different operator insertions in the core element of the synthetic minimal yeast promoter, four4 different insertion schemes were designed as follows (Figure S3 and Table S14): (1) directly downstream of the TATA box (TATA_OP), (2) 1 bp before the TATA box, (3) 1 bp before and directly downstream of the TATA box, and (4) directly downstream of the N30 sequence (N30_OP). We synthesized plasmids pKMV-FP1, pKMV-FP2, pKMV-FP3, and pKMV-FP4 on the basis of the UAS-FEC-core1 promoter,[15] which would produce the same sticky ends as the standardized promoters with *Bsm*BI digestion. Then, these synthetic promoters (SynPRO) were, respectively, assembled into the POT3 vector with HCKan_O_YPet and HCKan_T_tPKG1. The assembly reaction systems and the reaction program settings were similar to the POTB assembly. Then, the assembled transcription unit, SynPRO-YPet-tPGK1, was reversed in the POT3 vector to construct Assembled-Rev-POT3-ccdB-FP1-YPet-tPGK1, Assembled-Rev-POT3-ccdB-FP2-YPet-tPGK1, Assembled-Rev-POT3-ccdB-FP3-YPet-tPGK1, and Assembled-Rev-POT3-ccdB-FP4-YPet-tPGK1 by Gibson assembly. In these four plasmids, the UAS-FEC sequences were replaced by the *ccdB* gene with *Bsm*BI recognition sites at both sides, which produced the compatible sticky ends for UAS library insertion. Synthesis of the CcdB protein in the absence of CcdA protein leads to the induction of the SOS pathway and cell death,[25] which serves as a positive-selection marker.

Step 3: Reporter Module Assembly—UAS Assembly. As UASs significantly affect the reporter promoter strength, six 10 bp minimal UAS sequences (UAS_A, UAS_B, UAS_C, UAS_D, UAS_E, UAS_F) isolated through the library-based sorting and selection were linked in tandem to establish high-strength, short promoters in yeast, which made $6 \times 6 \times 6 = 216$ possible UAS designs (Figure S4a). Overall, 216 pairs of oligonucleotides with a 38 bp length were synthesized and each pair was annealed to make a UAS linker (Figures S4b and 4c). These linkers were first inserted into the HCKan_U fragment to construct the HCKan_UAS library (Table S15), which could release the UAS fragments with *Bsm*BI digestion. Then, the UAS sequences in the HCKan_UAS library were transferred into the ccdB-YPet-tPGK1 vector to assemble UAS-YPet-tPGK1 libraries (Table S16).

Step 4: Metabolite Biosensor Assembly—Barcode Assembly. The fourth step was to assemble the sensing module HC-AcV-POTB vectors constructed in step 1 and the reporter module in the UAS-YPet-tPGK1 libraries constructed in step 3 to build the metabolite biosensor and realize the assembly of molecular barcodes at the same time (Figure S5a). The HC-AcV-POTB vectors and UAS-YPet-tPGK1 libraries were digested with *Bpi*I and *Bsa*I, respectively (Table S17). The targeted bands in Figures S5b and 5c were recycled and then linked to construct 6 HC-AcV-POT-BC1-BC2-UAS-YPet-tPGK1 plasmid libraries (Table S18).

Step 5: Transformation of the Metabolite Biosensors into Yeast Cells. After the six HC-AcV-POT-BC1-BC2-UAS-YPet-tPGK1 plasmid libraries were successfully constructed, the genetic circuits on the HC-AcV vector were transferred to the shuttle plasmid AC reporter (Figure S6), which carried a URA3 selection marker. The AC-reporter plasmid harbored the pTEF2-mCherry-tADH1 transcription unit, which served as a control to normalize the cell-to-cell variability in gene expression (extrinsic noise). In addition, the AC reporter was a low copy plasmid and thus had smaller copy number differences between cells, which could further reduce the impact of extrinsic noise. At the same time, two *Bsa*I restriction sites were designed on the AC-reporter plasmid, which could be used for the insertion of the genetic circuits on the HC-AcV vector. The *ccdB* gene was inserted between the *Bsa*I restriction sites to improve assembly efficiency and reduce the impact of plasmid vectors. Furthermore, the AC-reporter plasmid and HC-AcV-POT-BC1-BC2-UAS-YPet-tPGK1 libraries were digested with *Bsa*I and *Bsm*BI, respectively (Table S19), to recover the target fragments as shown in Figure S12. Particularly, the plasmids that have not been successfully inserted in step 4 were removed after recycling the target bands of the HC-AcV-POT-BC1-BC2-UAS-YPet-tPGK1 plasmid library by agarose gel, which further improved the accuracy of the assembly (Figure S12). The finally constructed plasmid library was transformed into BY4700.

FACS Sorting Experiment. The glycerol stock of each separated library was inoculated in 14 mL high-clarity polypropylene (PP) round-bottom test tubes containing 4 mL of an SC-Ura medium and shaken at 220 rpm and 30 °C for 2 h. Then, each incubated stock was inoculated in 15 mL of SC-Ura media with an initial OD_{600} of 0.1 and divided into two 48-well deep well plates with 1 mL of culture in each well, as shown in Figure S13. After 12 h culture at 30 °C and 250 rpm, 2 μL of six distinct concentrations of cerulenin (0.5, 1.0, 1.5, 2.5, 4 mg/mL) was added to the corresponding wells. After another 12 h culture, the six separated libraries treated with the same cerulenin concentration were mixed and chilled on ice immediately. Cells in each sample were collected by centrifugation (4 °C; 8000*g* for 10 min) and resuspended in prechilled PBS to an OD_{600} of 2 to get Lib_0A&0B, Lib_1A&1B, Lib_2A&2B, Lib_3A&3B, Lib_5A&5B, and Lib_8A&8B. Then, all samples were kept on ice. BY4700 cultured in YPD media was used as a negative control. BY4700/POT1-pTEF2-mCherry-tADH1 and BY4700/POT1-pCYC1-YPet-tPGK1 cultured in SC-Ura media were used as positive controls for mCherry and YPet, respectively. BY4700/pGPD-TATA_OP-UAS_BBC (FapR-*fapO*-based malonyl-CoA sensor) cultured in SC-Ura media was used as a dual-positive control for both mCherry and YPet. These four control samples were also collected by centrifugation (4 °C; 8000*g* for 10 min) and resuspended in prechilled PBS to an OD_{600} of 2.

All cell sorting experiments were performed on an FACSAria SORP (BD Biosciences). Gating (P1) based on FSC-area and

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

SSC-area was performed to exclude noncell particles using the blank BY4700 strain. The fluorescence background noises for two colors were also calibrated using the blank BY4700 strain lacking YPet and mCherry expression. The resulting double-positive area in the panel of FITC-area (YPet) against PE-Texas Red-area (mCherry), which was determined by the positive control BY4700/POT1-pTEF2-mCherry-tADH1 and BY4700/POT1-pCYC1-YPet-tPGK1, was named after Q2. After all preparations, the dual-positive control BY4700/pGPD-TATA_OP-UAS_BBC was analyzed by cytometry, unveiling its cell distribution contour in Q2, which was a long ovate shape in the densest region. The major axis of this ovate was roughly parallel to the diagonal line of the log−log space, which divided the main body of the cells into two parts with approximately equal cell counts. Then, we set eight gates, P2−P9, in Q2, and the boundaries of P2−P3, P3−P4, P4−P5, P5−P6, P6−P7, P7−P8, and P8−P9 were set parallelly to the major axis. Then, the cell distribution area of the trackable library was sliced into eight similar narrow log-spaced strips, resulting in eight bins uniformly spanning Q2 (Figure S14).

In the main sorting process, 12 samples (trackable FapR-*fapO*-based malonyl-CoA biosensor treated with six distinct concentrations of cerulenin, Lib_0A&0B, Lib_1A&1B, Lib_2A&2B, Lib_3A&3B, Lib_5A&5B, Lib_8A&8B in two biological replicates) were individually sorted into the eight bins via a four-way purity mode. To eliminate the conflict events between two adjacent bins, P2, P4, P6, and P8 were simultaneously sorted in one run and so were P3, P5, P7, and P9. During the sorting process, the cell flow rate was kept at ∼10 000 events/s to keep the sorting efficiency at about 90%. In each bin, 10^5 to 5×10^5 cells were collected according to the actual cell density separately (Table S20).

NGS Library Preparation. The sorted cells were collected in 96 (12 samples × 8 bins) 5 mL sterile round-bottom polystyrene test tubes (BD Falcon), each containing 0.5 mL of 2× SC-Ura media. The collected cells were transferred into 14 mL high-clarity polypropylene (PP) round-bottom test tubes and fresh 2× SC-Ura media were added to make a 5 mL culture for each binned sample. All samples were grown at 30 °C and 220 rpm for 16 h and then collected for plasmid library extraction. The barcode region of each library was fetched through PCR amplification (Q5, 98 °C for 30 s, 30 cycles [98 °C, 10 s; 50 °C, 30 s; 72 °C, 6 s], 72 °C for 2 min) using the following pair of primers: NGS01&02 (P2), NGS03&04 (P3), NGS05&06 (P4), NGS07&08 (P5), NGS09&10 (P6), NGS11&12 (P7), NGS13&14 (P8), and NGS15&16 (P9). In a 50 μL reaction, 12.5 μL of the plasmid library was added as a PCR template. PCR conditions including annealing temperature and PCR cycles were optimized to alleviate biased amplification to different sequences in the library.[26] The resulting 96 PCR products were first analyzed by electrophoresis to ensure an expected 131 bp band (containing UAS sequences, BC1 and BC2) without any other nonspecific band corresponding to the PCR side product. Subsequently, the eight PCR products for eight bins from each sample were mixed and then delivered to Genewiz for sequencing library construction and NGS analysis. Sequencing was carried out using a 2 × 150 paired-end configuration. Approximately, 30 million reads were collected for each library (Table S21).

NGS Data Processing. First, the raw NGS data were demultiplexed and the adapter region was removed by cutadapt[27] to produce clean data for each sequencing library. Second, the clean data were merged into single-end data by FLASH script[28] via removing those two ends that suffered from mismatches (min-olap: 56, max-olap: 150). The length of about 85% merged reads was 131 bp, in line with expectations. Subsequently, the fastq file obtained by FLASH merging was stripped into different fasta files according to the corresponding gate barcode (Table S22) sequences contained at both ends. Then, the stripped fasta files were mapped to the Combination_seq.fasta, which contained 5184 combination names and their corresponding barcode sequences, by seqmap1.0.13.[29] Each combination was named as TF_promoter (*pGPD*, *pENO2*, *pHSP12*, *pEXG1*, *pCYC1*, *pULI1*)−operator insertion scheme (TATA_OP, OP_TATA, OP_TATA_OP, N30_OP)-UAS_link in tandem (AAA−FFF). For combination s_k, its number of reads in sorting gate P_j, r_{j,s_k}, was obtained through this comparison process (#mismatch = 0), and the total number of reads obtained in gate P_j was r_j.

Deriving the μ Value for Each Combination in a Trackable Manner from FACS-seq Data. The expression of genes in cells involves complex biochemical processes. Even if the genotypes are the same, the expression level of genes presents a log-normal distribution.[16,17] Therefore, the distribution of each combination across all bins should theoretically follow a log-normal distribution, as observed experimentally (Figure S1b). Here, we use the $\log_{10}$ ratio of YPet/mCherry expression, termed as the μ value, to characterize the biosensor response. At first, the lower and upper boundary values of sorting gate P_j, $A_{b,j}$, and $A_{u,j}$, were determined (for bin P2, $A_{u,2}$ was the 95th quantile of all cells in the P2; for bin P9, $A_{b,9}$ was the 5th quantile of all cells in the P9). The median values of the fluorescence intensity I_j of each bin were calculated as $I_j = (A_{b,j} + A_{u,j})/2$ (Table S23). According to the NGS data, we actually observed r_{j,s_k} read counts for cells carrying combination s_k in bin j. Hence, assuming an unbiased NGS quantification process, the actual ratio of cells carrying combination s_k in bin j, should be $R_{j,s_k} = r_{j,s_k}/r_j$. The possibility of cells carrying combination s_k to be sorted into bin j was

$$p_{j,s_k} = \frac{R_{j,s_k} \times C_j}{\sum_{P2}^{P9} R_{j,s_k} \times C_j}$$

where C_j is the ratio of cells sorted into bin j against all cells in Q2 in one sorting experiment (Table S24). Then, the μ value of combination s_k was derived as the weighted average using the median of $\log_{10}$(YPet/mCherry) of each bin

$$\mu_{s_k} = \sum_{P2}^{P9} I_j \times p_{j,s_k}$$

Individual Characterization. Colonies were picked from the plates and inoculated into 48-well deep well plates with 1 mL of SC-Ura medium in each well. After 24 h culture at 30 °C and 250 rpm, 2 μL of six distinct concentrations of cerulenin (0.5, 1.0, 1.5, 2.5, 4 mg/mL) was added to the corresponding well. After another 12 h culture, cells in each well were collected by centrifugation (4 °C; 8000*g* for 10 min) and resuspended in prechilled PBS to an OD_{600} of 2. Then, all samples were kept on ice. The fluorescence intensities of the cells were characterized on an LSRFortessa (BD Biosciences), which shared a similar protocol with the FACS sorting experiment. For each sample, 10 000 cells in the Q2 area were analyzed.

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.1c00595.

Additional experimental details, materials, methods, and cytometry analysis results using pGPD-TATA_OP-UAS_BBC; description of the trackable assembly workflow; predicted μ value distribution histograms of different operator insertion schemes, different UAS, and different promoters of *fapR* when no cerulenin was added; predicted μ value distribution histograms in different cerulenin concentrations; example of the trackable assembly of the AND logic biosensor; validation of target bands by agarose gel electrophoresis; schematic diagram of the distribution of FACS library samples in 48-well deep well plates; FACS sorting experiment of Lib_0A shown in the cytometry panel (Figures S1−S14); diversity analysis of the trackable assembly of biosensors; sequencing results of the colonies sampled from the mixed library and the separated library; distribution of 2632 biosensors characterized by FACS-seq; Gini coefficients of the distribution of Lib0A; strains, plasmids, primers, and other oligonucleotides used in this work; prefix and suffix sequences of the genetic elements used in the POTB assembly; description of promoters driving the expression of FapR; reaction systems and program settings used in the trackable assembly; information of the barcodes used in this work; median of the fluorescence intensities of each bin; number and the ratio of cells sorted into each bin; and data qualities of NGS libraries (Tables S1−S6 and S8−S24) (PDF)

Cost calculation for tuning the dose−response curve in a massively parallel manner and an arrayed manner (Table S7) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Chong Zhang − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China; Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China;* orcid.org/0000-0001-9609-8855; Email: chongzhang@tsinghua.edu.cn

Authors

Yikang Zhou − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China*

Yaomeng Yuan − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China*

Yinan Wu − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China*

Lu Li − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China;* orcid.org/0000-0001-8957-5246

Aysha Jameel − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China*

Xin-Hui Xing − *MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China; Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.1c00595

Author Contributions

[§]Y.Z. and Y.Y. contributed equally to this work. Y.Z. conducted the experiments and analyzed the data and Y.Y. and Y.Z. drafted the manuscript. All authors revised and approved the manuscript.

Notes

The authors declare no competing financial interest.

NGS raw data of FACS-seq can be accessed from the NCBI Bioproject via accession number BioProject PRJNA782357. The homemade python script used to process FACS-seq raw data can be accessed on https://github.com/yuanym19/FACS_seq_FapR_Based_Biosensor_Fine_Tuning. Any other data or materials related to this work are available upon request from the corresponding author.

■ ACKNOWLEDGMENTS

The authors would like to thank Bin Yu at the Center of Biomedical Analysis, Tsinghua University, for the support in the FACS experiment. This work was supported by the National Key Research and Development Program of China (2019YFA0904800) and the National Natural Science Foundation of China (U2032210 and U1932136).

■ REFERENCES

(1) Liang, W. F.; Cui, L. Y.; Cui, J. Y.; Yu, K. W.; Yang, S.; Wang, T. M.; Guan, C. G.; Zhang, C.; Xing, X. H. Biosensor-assisted transcriptional regulator engineering for Methylobacterium extorquens AM1 to improve mevalonate synthesis by increasing the acetyl-CoA supply. *Metab. Eng.* **2017**, *39*, 159−168.

(2) Hanko, E. K. R.; Minton, N. P.; Malys, N. A Transcription Factor-Based Biosensor for Detection of Itaconic Acid. *ACS Synth. Biol.* **2018**, *7*, 1436−1446.

(3) Liang, C.; Zhang, X.; Wu, J.; Mu, S.; Wu, Z.; Jin, J. M.; Tang, S. Y. Dynamic control of toxic natural product biosynthesis by an artificial regulatory circuit. *Metab. Eng.* **2020**, *57*, 239−246.

(4) Xu, P.; Li, L.; Zhang, F.; Stephanopoulos, G.; Koffas, M. Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 11299−11304.

(5) Riglar, D. T.; Giessen, T. W.; Baym, M.; Kerns, S. J.; Niederhuber, M. J.; Bronson, R. T.; Kotula, J. W.; Gerber, G. K.; Way, J. C.; Silver, P. A. Engineered bacteria can function in the mammalian gut long-term as live diagnostics of inflammation. *Nat. Biotechnol.* **2017**, *35*, 653−658.

(6) Engler, C.; Kandzia, R.; Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **2008**, *3*, No. e3647.

(7) Smanski, M. J.; Bhatia, S.; Zhao, D.; Park, Y. J.; Woodruff, L. B. A.; Giannoukos, G.; Ciulla, D.; Busby, M.; Calderon, J.; Nicol, R.; Gordon, D. B.; Densmore, D.; Voigt, C. A. Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* **2014**, *32*, 1241−1249.

(8) Gibson, D. G.; Young, L.; Chuang, R. Y.; Venter, J. C.; Hutchison, C. A.; Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **2009**, *6*, 343−345.
(9) Cohen, S. N.; Chang, A. C. Y.; Boyer, H. W.; Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 3240−3244.
(10) Shetty, R. P.; Endy, D.; Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.* **2008**, *2*, No. 5.
(11) Anderson, J. C.; Dueber, J. E.; Leguia, M.; Wu, G. C.; Arkin, A. P.; Keasling, J. D.; et al. BglBricks: A flexible standard for biological part assembly. *J. Biol. Eng.* **2010**, *4*, No. 1.
(12) Townshend, B.; Kennedy, A. B.; Xiang, J. S.; Smolke, C. D. High-throughput cellular RNA device engineering. *Nat. Methods* **2015**, *12*, 989−994.
(13) Adams, R. M.; Mora, T.; Walczak, A. M.; Kinney, J. B. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife* **2016**, *5*, No. e23156.
(14) Wong, A. S. L.; Choi, G. C. G.; Cheng, A. A.; Purcell, O.; Lu, T. K. Massively parallel high-order combinatorial genetics in human cells. *Nat. Biotechnol.* **2015**, *33*, 952−961.
(15) Redden, H.; Alper, H. S. The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.* **2015**, *6*, No. 7810.
(16) Elowitz, M. B.; Levine, A. J.; Siggia, E. D.; Swain, P. S. Stochastic gene expression in a single cell. *Science* **2002**, *297*, 1183−1186.
(17) Raser, J. M.; O'shea, E. K. Control of Stochasticity in Eukaryotic Gene Expression. *Science* **2004**, *304*, 1811−1814.
(18) Liang, J. C.; Chang, A. L.; Kennedy, A. B.; Smolke, C. D. A high-throughput, quantitative cell-based screen for efficient tailoring of RNA device activity. *Nucleic Acids Res.* **2012**, *40*, No. e154.
(19) Cheng, A. A.; Ding, H.; Lu, T. K. Enhanced killing of antibiotic-resistant bacteria enabled by massively parallel combinatorial genetics. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 12462−12467.
(20) Xu, P.; Wang, W.; Li, L.; Bhan, N.; Zhang, F.; Koffas, M. A. G. Design and kinetic analysis of a hybrid promoter-regulator system for malonyl-CoA sensing in Escherichia coli. *ACS Chem. Biol.* **2014**, *9*, 451−458.
(21) Chen, X.; Yang, X.; Shen, Y.; Hou, J.; Bao, X. Screening phosphorylation site mutations in yeast acetyl-CoA carboxylase using malonyl-CoA sensor to improve malonyl-CoA-derived product. *Front. Microbiol.* **2018**, *9*, No. 47.
(22) Qiu, C.; Zhai, H.; Hou, J. Biosensors design in yeast and applications in metabolic engineering. *FEMS Yeast Res.* **2019**, *19*, No. foz082.
(23) Li, S.; Si, T.; Wang, M.; Zhao, H. Development of a Synthetic Malonyl-CoA Sensor in Saccharomyces cerevisiae for Intracellular Metabolite Monitoring and Genetic Screening. *ACS Synth. Biol.* **2015**, *4*, 1308−1315.
(24) Guo, Y.; Dong, J.; Zhou, T.; Auxillos, J.; Li, T.; Zhang, W.; Wang, L.; Shen, Y.; Luo, Y.; Zheng, Y.; Lin, J.; Chen, G. Q.; Wu, Q.; Cai, Y.; Dai, J. YeastFab: The design and construction of standard biological parts for metabolic engineering in Saccharomyces cerevisiae. *Nucleic Acids Res.* **2015**, *43*, No. e88.
(25) Van Melderen, L.; Bernard, P.; Couturier, M. Lon-dependent proteolysis of CcdA is the key control for activation of CcdB in plasmid-free segregant bacteria. *Mol. Microbiol.* **1994**, *11*, 1151−1157.
(26) Gohl, D. M.; Vangay, P.; Garbe, J.; MacLean, A.; Hauge, A.; Becker, A.; Gould, T. J.; Clayton, J. B.; Johnson, T. J.; Hunter, R.; Knights, D.; Beckman, K. B. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **2016**, *34*, 942−949.
(27) Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **2011**, *17*, 10−12.
(28) Magoč, T.; Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **2011**, *27*, 2957−2963.
(29) Jiang, H.; Wong, W. H. SeqMap: Mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **2008**, *24*, 2395−2396.

https://doi.org/10.1021/acssynbio.1c00595
ACS Synth. Biol. 2022, 11, 977−989