Data-Driven Prediction of CRISPR-Based Transcription Regulation for Programmable Control of Metabolic Flux

Jiayuan Sheng†, Weihua Guo†, Christine Ash, Brendan Freitas, Mitchell Paoletti, and Xueyang Feng*

Department of Biological Systems Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

† WG and JS are equally contributed.

* To whom correspondence should be addressed. X.F: Phone: (540) 231-2974. E-mail: xueyang@vt.edu.

Abstract

Multiplex and multi-directional control of metabolic pathways is crucial for metabolic engineering to improve product yield of fuels, chemicals, and pharmaceuticals. To achieve this goal, artificial transcriptional regulators such as CRISPR-based transcription regulators have been developed to specifically activate or repress genes of interest. Here, we found that by deploying guide RNAs to target on DNA sites at different locations of genetic cassettes, we could use just one synthetic CRISPR-based transcriptional regulator to simultaneously activate and repress gene expressions. By using the pairwise datasets of guide RNAs and gene expressions, we developed a data-driven predictive model to rationally design this system for fine-tuning expression of target genes. We demonstrated that this system could achieve programmable control of metabolic fluxes when using yeast to produce versatile chemicals. We anticipate that this master CRISPR-based transcription regulator will be a valuable addition to the synthetic biology toolkit for metabolic engineering, speeding up the "design-build-test" cycle in industrial biomanufacturing as well as generating new biological insights on the fates of eukaryotic cells.

Introduction

Metabolic engineering has proven to be tremendously important for sustainable production of fuels[1,2], chemicals[3,4], and pharmaceuticals[5,6]. One of the critical steps in metabolic engineering is reprogramming metabolic fluxes in host cells to optimize the fermentation performance such as product yield[7,8]. To achieve this goal, several enzymes need to be activated while in the meantime others need to be repressed[9]. The multiplex and multi-directional control of enzyme expressions is largely executed through transcriptional regulation[9-12]. Recently, the type-II clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 from *Streptococcus pyogenes* (*Sp*)[13-15] has been repurposed to be a master transcriptional regulator that could activate or repress multiple genes[16-18]. By further extending the guide RNAs of SpCas9 to include effector protein recruitment sites and expressing the effector proteins in host cells[18,19], a synthetic CRISPR-based transcriptional regulator was developed to simultaneously program the expressions of multiple genes at multiple directions (i.e., both activation and repression).

Although demonstrating great promises in controlling metabolic fluxes, the current CRISPR-based transcription regulation faces several challenges when being applied for metabolic engineering. First, it relies on a panel of well-characterized genetic parts (e.g., RNA-binding proteins) to achieve the transcriptional regulation[9,20]. However, the rareness of such genetic parts often limits the application of current CRISPR-based transcription regulator in a metabolic network. Second, it requires the co-expression of effector proteins to activate or repress the target genes[9,20]. Therefore, the utility of current CRISPR-based transcription regulator in metabolic engineering is mitigated by the metabolic costs associated with protein expressions[9,20]. To address these limitations, an ideal type of CRISPR-based transcription regulation should meet two criteria: generally applicable in any gene of interest, and requiring

minimal protein expression to achieve multi-directional gene regulation.

In this study, we have developed a data-driven approach that enables the rational design of a new type of CRISPR-based transcription regulation. This CRISPR-based transcription regulator uses only a fused protein of a nuclease-deficient Cas9 (dCas9) and an effector (VP64) to achieve multi-directional and multiplex gene regulation, which eliminates the metabolic costs associated with the expression of effector proteins in previous studies[9,20]. We found that the deployment of guide RNAs was the key factor that determined the regulatory effects of our CRISPR-based transcription regulator. We used a data-driven approach to provide accurate and target-oriented guidance on designing the guide RNAs. As we showed in our results, this approach could be applied for any gene of interest. Finally, using this system, we demonstrated a highly programmable control of metabolic fluxes when using yeast to produce versatile chemicals.

Results

Gene activation and repression by using a CRISPR-based transcriptional regulator. The CRISPR-based transcriptional regulator is composed of a codon-optimized, catalytically dead SpCa9 (i.e., dCas9) that is fused with four tandem copies of Herpes Simplex Viral Protein 16 (VP64, a commonly used eukaryotic transcription activator domain). The similar molecular design was previously reported to be able to activate gene expression in yeast[18] and mammalian cells[18]. In brief, it was found that when the dCas9-VP64 regulator was positioned in the correct sites of promoter, the target gene could be activated by the VP64. However, we hypothesized that the effects of dCas9-VP64 could be diverse, i.e., both activation and repression could be achieved by using this master regulator (Fig. 1A). For example, the transcription initiation could

be blocked when dCas9-VP64 is deployed to the transcription starting sites (TSS). The transcription elongation could also be inhibited when dCas9-VP64 is deployed to the open reading frame (ORF). If the hypothesis stands, we could then repurpose dCas9-VP64 as a universal regulator to activate and repress gene expressions at the same time.

To validate our hypothesis, we designed experiments by selecting 138 sites that can be targeted by dCas9-VP64 in four synthetic genetic cassettes (Fig. 1B): GFP under TEF1p promoter, mCherry under TPI1p promoter, Sapphire under PGK1p promoter, and Venus under PDC1p promoter. We created a library of guide RNAs and co-expressed them with the dCas9-VP64 and the target genetic cassette. For each of the tests, we measured the fluorescence of the reporter proteins and compared it to a control test in which dCas9-VP64 and the genetic cassette were expressed without guide RNA. As summarized in Fig. 1B and Fig. 1C, expression of reporter genes could indeed be programmed to be either up- or down-regulated (cutoff fold-change set as two-fold, $p<0.05$) when positioning the guide RNA at different sites on the promoter or ORF. The dynamic range of transcriptional regulation via dCas9-VP64 varied for different genetic elements, with the largest dynamic range achieving 13.8-fold (from -1.14 to 2.65 of log2 gene fold-change) in expression of TPI1p-mCherry cassette and the smallest $\log_2$ fold change dynamic range achieving 11.6-fold (from -1.82 to 1.71 of log2 gene fold-change) in expression of PGK1p-Venus cassette. We next compared the effects on transcriptional regulation when guide RNAs were positioned in the promoter region to that were positioned in the gene region (Fig. 1D). A significant difference ($p<0.01$) was revealed: while gene expression was in general down-regulated when guide RNAs were positioned in the gene region, most of the guide RNAs (94.2%) that were positioned in the promoter region led to gene up-regulation or no effect. We also evaluated whether different PAM sites (i.e., 3'AGG, 3'TGG, 3'CGG, 3'GGG) could

bias the transcriptional regulation (Fig. 1E). We found that most of the PAM sites (3'AGG, 3'TGG, 3'CGG) have no bias, but when guide RNAs were targeted on 3'GGG sites, the expression of selected genes tends to be more up-regulated than other PAM sites (p<0.05). Such difference, as we discussed in the section below, may be attributed to the larger percentage of guanosine in 3'GGG sites. Finally, we evaluated the metabolic costs of dCas9-VP64 by comparing the growth rates of yeast between the ones subject to transcriptional regulation (i.e., with guide RNAs) and the ones that were not. As shown in Fig. S1, 131 out of the 138 tests showed no significant difference (p>0.05) on cell growth rate when being compared to that of the control strain, indicating a minimal metabolic burden when using only one synthetic protein for transcriptional regulation.

Data-driven model of transcriptional regulation by using dCas9-VP64. To determine the rule underlying transcriptional regulation by dCas9-VP64, we solicited nine design parameters on nucleotide stability, sequence of the target genetic element, PAM site location, and protein-DNA structure. These design parameters were chosen based on previous studies on the activity of SpCas9 [21,22]. Next, we correlated these design parameters with transcriptional regulation (Fig. S2A), and calculated the Pearson's correlation coefficients (PCC). The top 3 correlated design parameters were location, GC content (GC%), and PAM site of GGG. Next, we aimed to develop a predictive model that could use the design parameters to describe the effects of guide RNAs positioning on gene expressions. Our first attempt is a linear regression model, which utilized all the nine design parameters (i.e., GC%, location, number of base G, number of base A, ΔG, AGG, TGG, CGG, and GGG) to simulate the corresponding fold changes of gene expressions as we collected from the four synthetic genetic cassettes (Fig. S2B). However, the fitting of the linear regression model was very bad, as demonstrated by the low PCC (0.41)

between observed and simulated fold changes (Fig. S2B). This clearly indicated that the biomolecular interactions of RNA-protein and DNA-protein in the dCas9-VP64 system were highly nonlinear, which cannot be captured by the simple linear regression model.

We then used machine learning to derive an empirical model to quantitatively predict the nonlinear correlation between design parameters and transcriptional regulation by dCas9-VP64. As a data-driven modeling approach, machine learning is advantageous than linear model in solving complex problems in two aspects[23-25]: requiring no *a priori* knowledge of the system, and capable of resolving complex systems with high non-linearity and multi-dimensionality. In this study, we used the pairwise data of the design parameters of guide RNAs and the corresponding fold-change of gene expressions to train the computer for developing a mathematical model that could accurately predict the causal effects of inputs (i.e., regulated gene expression in this study). We used decision tree method to build a machine-learning algorithm and adopted ten-fold cross validation to evaluate the prediction accuracy of our model. The model construction and model evaluation were conducted by following a toolkit developed in MATLAB™ (i.e., "*fitrtree*" and "*predict*" in Statistics and Machine Learning Toolbox), which automatically adjust the nodes and connections of the decision tree to optimize the fitting[26,27]. Using the 138 pairwise data collected from the synthetic genetic cassettes (i.e., TEF1p-GFP, TPI1p-mCherry, PDC1p-Sapphire, and PGK1p-Venus), we found that the prediction accuracy was dramatically improved with PCC reaching 0.80 (Fig. 2B) for overall prediction of gene regulation and 0.72~0.93 for predicting gene regulation of individual genetic cassettes (Fig. 2C). Also, we found that the prediction accuracy of our machine-learning algorithm was improved with the enlargement of data size (Fig. 2D). For example, when 20 pairwise datasets were chosen, the PCC was merely 0.66. However, when the number of pairwise datasets reached 80, the PCC

was improved to 0.78. This demonstrated the unique advantage of data-driven algorithm, i.e., increased prediction accuracy with more data.

To further test if our machine-learning algorithm could be generally applied for predicting the transcriptional regulation of other genes, we designed another synthetic genetic cassette that expressed tdTomato under Eno2p promoter. We used the machine-learning algorithm to predict the regulated gene expressions when guide RNAs of dCas9-VP64 were positioned at different locations of Eno2p-tdTomate cassette, followed by constructing the genetic cassettes and conducting experimental measurements. Of the 99 guide RNAs tested, our machine-learning algorithm could achieve similarly high prediction accuracy with PCC between the predicted and the measured gene expressions reaching 0.87 (Fig. 2E). This success demonstrated that our data-driven model could be generally applicable to guide the biomolecular design of dCas9-VP64 system to achieve customized gene regulation.

We also packaged our machine-learning algorithm into an open-source toolbox (Fig. S3 and supplementary software), CRISTINES (CRISPR-Cas9 Transcriptional Inactivation aNd Elevation System), which is a MATLAB-based toolbox and free for downloading at https://sites.google.com/a/vt.edu/biomolecular-engineering-lab/software. CRISTINES is able to analyze the input DNA sequences, identify the design parameters of the dCas9-VP64 system, and use the embedded decision-tree algorithm to provide the top five guide RNAs that would lead to the strongest up-regulation and down-regulations, respectively. We anticipated that CRISTINES could help biologists worldwide to customize their design based on the target gene of interest.

Design dCas9-VP64 to reprogram metabolic fluxes in yeast. We next applied dCas9-VP64 system in yeast metabolic engineering to test if the metabolic fluxes could be flexibly

reprogrammed using our CRISPR-based transcription regulator. We chose the highly branched bacterial violacein biosynthetic pathway as our model pathway[28] (Fig. 3A), which uses five enzymes (VioA, VioB, VioE, VioD, and VioC) to produce four high-value products (violacein, proviolacein, deoxyviolacein, and prodeoxyviolacein). By controlling the expression levels of the five enzymes, the metabolic fluxes flowing into different branch pathways could be varied and thus leading to different yield of the four products. In this study, we reconstituted the violacein pathway in yeast by expressing the five enzymes under constitutive promoters (i.e., TEF1p, PGK1p, ENO2p, TPI1p, and PDC1p). According to our data-driven model CRISTINES, we could computationally predict the gene expression and thus predict the metabolic fluxes in the violacein pathway. Here, we designed three guide RNAs for four genes in the violacein pathway (VioA, VioE, VioD, and VioC). These three guide RNAs targeted on different promoter sites and were predicted to result in high, medium and low expression of each target gene, respectively. Correspondingly, the functional output states of the violacein pathway were predicted to vary. For example, up-regulating VioA and down-regulating VioD would increase the fluxes into deoxyviolacein and prodeoxyviolacein, but decrease the fluxes into violacein and proviolacein. To validate our predictions on metabolic flux reprogramming, we co-expressed the dCas9-VP64 system with the violacein pathway, and used various guide RNAs to fine tune gene expressions. For each of the tests, we measured the titer of violacein, proviolacein, deoxyviolacein, and prodeoxyviolacein produced by yeast. As shown in Fig. 3B and 3C, our prediction fit well with the experimental measurements, with PCC reaching 0.84. We noticed some of the predictions on metabolic fluxes were not as good as we expected. This could be attributed to the posttranscriptional regulation of the five enzymes in the violacein pathway. Overall, we demonstrated that our master CRISPR-based transcription regulator was indeed able

to program metabolic fluxes. More importantly, our data-driven algorithm allows users to design metabolic pathways with deterministic fates *in silico*.

Discussion

Using data-driven approach to investigate biomolecular interactions of CRISPR-based systems has recently been showcased in several studies, such as rational design of guide RNAs for maximizing editing activity and minimizing off-target effects of SpCas9[29,30]. This approach is advantageous compared to conventional deterministic models because it does not require *a priori* knowledge on the mechanisms of RNA-protein and DNA-protein interactions[31,32], which still remains largely unknown in spite of numerous studies on SpCas9 structures[33-37]. Also, because the biomolecular interactions among Cas9 protein and nucleic acids are highly nonlinear, the linear regression model cannot capture the essence of CRISPR-based transcription regulation. As shown in our results, a machine-learning method could overcome this issue and capture the nonlinearity of the model. Not only did we achieve high accuracy when predicting the regulatory effects of CRISPR-based transcription regulator, but we also demonstrated that this method was not specific to a few selected genes and could be generally applicable. Future work will determine if the results obtained from yeast could provide useful lessons in other eukaryotic systems such as mammalian cells.

We expect that our method will provide a valuable tool for metabolic engineering, especially yeast metabolic engineering at this stage. *S. cerevisiae* is a widely used industrial workhorse for producing a broad spectrum of chemicals that represents over quarter trillion dollars market[38]. The experimental and analytical approaches described here raise the possibility of genome-scale reprogramming of metabolic fluxes, which will dramatically speed up the

"design-build-test" cycle in industrial biomanufacturing[39]. We also expect our method could be used to rewire the fate of yeast cells, such as cell cycle, and thus generate new biological insights on the fundamentals of metabolic diseases, aging and apoptosis by using yeast as a disease model.

Methods

Strain and plasmid construction in *Saccharomyces cerevisiae*. dCas9 was codon-optimized for expression in *S. cerevisiae* and cloned into a pRS413 backbone under control of the GAL1 promoter. The RNA-guided transcription factors were built by fusing four repeats of the minimal domain of the herpes simplex viral protein 16 (VP16) to the C-terminus of dCas9 (dCas9_VP64)[18]. The reporter genes eGFP under the control of the TEF1p promoter, sapphire under PDC1p, mCherry under TPI1p, and venus under PGK1p were cloned into pRS416 plasmid by using the DNA assembler method[40,41]. The reporter plasmid for verification was built by cloning tdTomato under the control of ENO2p into pRS416 plasmid. To build gRNA-expressing plasmids, empty gRNA expressing vectors were first made by cloning the pRPR1 promoter (an RNA-polymerase-III-dependent promoter), the gRNA handle (flanked by HindIII and Xho1 sites), and the RPR1 terminator into the SacI and KpnI sites of the pRS425 plasmid. Sequences of the constructs that were used in this study were listed in Table S1. Strains constructed in this study were listed in Table S2.

Fluorescence Assays. To assess expression of the reporter constructs, yeast cells expressing different gRNAs (or no gRNA as control) were grown overnight (250 rpm, 30°C) in 3 mL SC medium supplemented with glucose with appropriate selection (three independent cultures for each sample). Ten microliters of these cultures were then transferred into fresh media, supplemented with galactose and grown for 20 h (250 rpm, 30°C) before analysis by plate

reader. The wave lengths of the different reporter genes are eGFP: λ_{ex} 488 nm, λ_{em} 507nm; Sapphire: λ_{ex} 399 nm, λ_{em} 511nm; Venus λ_{ex} 515 nm, λ_{em} 528nm; tdTomato λ_{ex} 554 nm, λ_{em} 581nm; mCherry λ_{ex} 587 nm, λ_{em} 610 nm. All of the fold-changes of the synthetic genetic cassettes, including the four cassettes for model training and the ENO2p-tdTomato cassette for model validation, were listed in Table S3 and Table S4.

Qualitative analysis of key parameters. The exponential fold changes of different gRNA designs have been separated into two categories, the promoter region (location < 0, n = 52) and the gene coding region (location > 0, n = 86). The unpaired two-tail t-test was used to calculate the significance of the fold changes between these two groups. For the analysis of the PAM type, the same t-test was used for each PAM types (n_{AGG} = 42, n_{TGG} = 69, n_{CGG} = 21, n_{GGG} = 6).

Modeling. The multiple linear regression model was implemented by the "*regress*" command in MATLAB. To evaluate the prediction power, the ten-fold cross validation was implemented for the linear model. The Pearson's correlation coefficient between observed and predicted fold changes was calculated by MATLAB. The binary regression decision tree model was developed by using "*fitrtree*" command in MATLAB with all the default setting of options. The details of the decision tree model were shown in Table S5. The ten-fold cross validation was implemented for this model with the same manner. A detailed tutorial of CRISTINES was included in the supplementary information.

Ten-fold cross validation. All the datasets from the four synthetic genetic cassettes (n = 151) were randomly divided into ten folds (nine folds with 15 datasets each and one-fold with 16 datasets). Nine of the ten folds were used as training datasets, and the one-fold remaining was used as validation datasets. By iteratively repeating the above process for ten times, all the folds

could be used as validation datasets. The Pearson's correlation coefficient between the observed fold changes and predicted fold changes of were calculated by MATLAB.

Analysis of products from violacein pathway. Yeast strains for violacein biosynthesis were constructed and product distributions were analyzed as described previously with minor modifications. The parent yeast strain for these experiments was BY4741. The five-gene cassette of violacein pathway was constructed using the DNA assembler method: VioA under TEF1p; VioB under PGK1p; VioC under ENO2p; VioD under TPI1p and VioE under PDC1p. Yeast strains with violacein pathway genes and the CRISPR system with constitutive dCas9 expression were grown in SC medium containing 5% galactose. After 3 days at 30 °C, approximately 2 mL of yeast cultures were harvested and the cells were collected and suspended in 250 μL of methanol, boiled at 95 °C for 15 minutes, and vortexed twice during the incubation. Solutions were centrifuged twice to remove cell debris, and the products from violacein pathway (i.e., violacein, proviolacein, deoxyviolacein, and prodeoxyviolacein) in the supernatant were analyzed by HPLC on an Agilent Rapid Resolution SB-C18 column as described previously, measuring absorbance at 565 nm[42].

Code availability. We claim that the code mentioned in this study is available within the article's Supplementary Information files (Supplementary Software) and at https://sites.google.com/a/vt.edu/biomolecular-engineering-lab/ as a MATLAB package file (MathWorks Inc.).

Data availability. All data generated or analyzed during this study are included in this published article and its Supplementary Information files.

Acknowledgement

We thank the Writing Center in Virginia Tech for improving the language of the paper. This study was supported by start-up fund (#175323) and the ICTAS Junior Faculty Award from Virginia Tech.

Author contributions

X.F. contributed to the initial idea of this project. J.S., C.A., and B.F. contributed to the molecular biology experiments and data screening. W.G. and X.F. contributed to the computational modeling and analysis. J.S. and W.G. contributed to the experimental validation and violacein pathway showcase. W.G. and M.P. contributed to the offline software development. J.S., W.G., and X.F. contributed to the manuscript preparation and revision.

Competing financial interests.

The authors do not have any conflicts of financial interests.

Supplementary Information

Figure S1. The OD_{600} fold change for all the strains used for screening. (A) OD_{600} fold-change for GFP screening set. (B) OD_{600} fold-change for Sapphire screening set. (C) OD_{600} fold-change for mCherry screening set. (D) OD_{600} fold-change for Venus screening set. The results showed that no significant metabolic burden could be detected compared with wild type strain. The strains marked with (*) indicated a significant inhibition of growth ($p<0.05$).

Figure S2. Linear model for predicting transcriptional regulation by dCas9-VP64. (A) Pearson's correlation coefficients between each guide RNA design parameters and the fold changes of gene expressions from the screening data of the four synthetic genetic cassettes. (B) Simulation accuracy of the linear model. (C) Validation of the linear model by comparing the simulated and

experimentally measured gene regulations on Eno2p-tdTomato cassette subject to dCas9-VP64 regulation.

Figure S3. Screenshot of CRISTINES software with a demo sequence.

Table S1. DNA sequences and plasmids used in this study.

Table S2. Strains used in this study.

Table S3. Key parameters of guide RNAs used in this study.

Table S4. Key parameters of guide RNAs used in validation experiments (Eno2p-tdTomato).

Table S5. Binary regression decision tree model.

Table S6. Measured and predicted concentrations of products from violacein pathway

Supplementary Software. CRISTINES toolbox.

Figures and figure legends

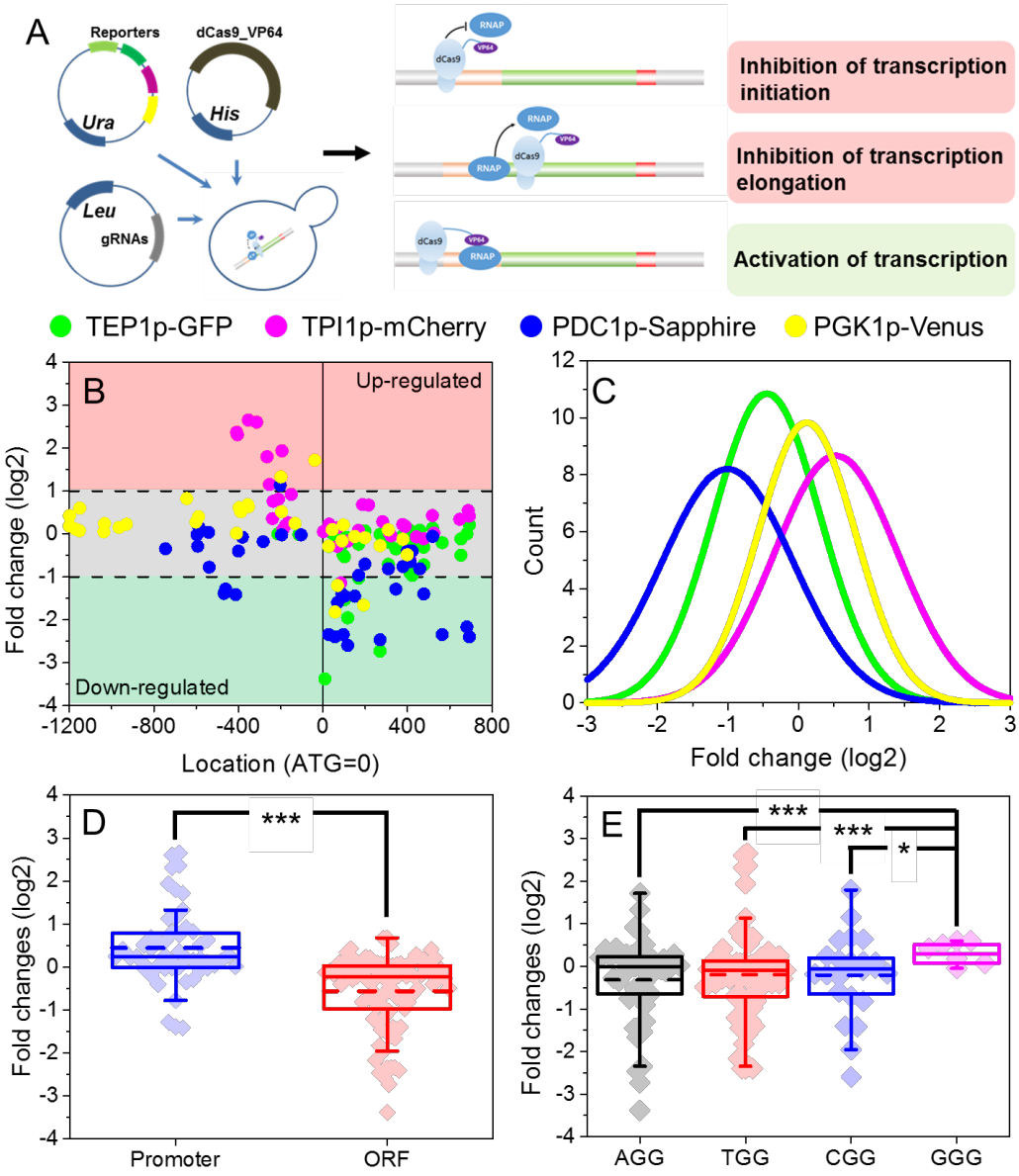

Figure 1. Multi-directional transcriptional regulation by dCas9-VP64. (A) Hypothesized mechanism of transcriptional regulation by dCas9-VP64 to achieve both gene activation and gene repression. (B) The measured fold changes of gene expressions from the four synthetic genetic cassettes based on the PAM position. (C) The distributions of fold changes of gene expressions. (D) Comparison of fold changes of gene expressions from two groups: PAM sites located in the promoter regions and PAM sites located in the ORF region. ***: $p < 0.01$. (E) Effects of different PAM sites on transcriptional regulation by dCas9-VP64. *: $p < 0.05$.

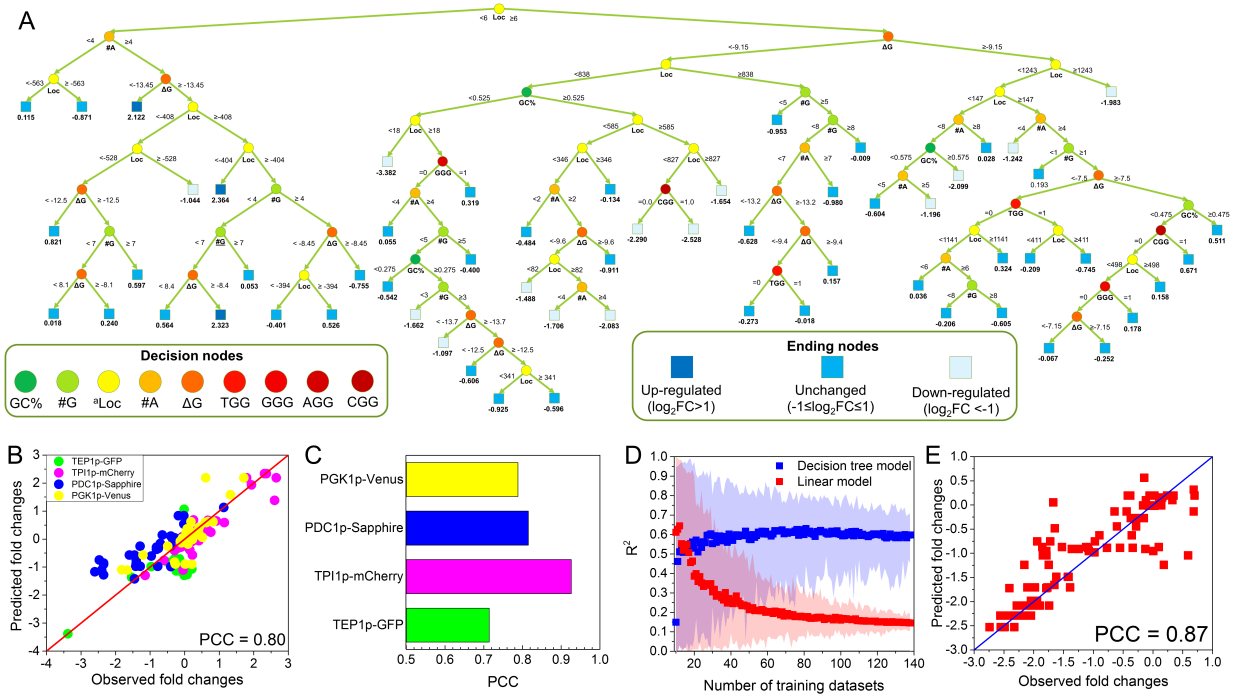

Figure 2. Data-driven model of transcriptional regulation by using dCas9-VP64. (A) Binary regression tree model trained with all the screening data from the four synthetic genetic cassettes. The regression tree model consisted of 58 decision nodes and used six design parameters of guide RNAs as input. (B) and (C) Prediction accuracy of the regression tree model from ten-fold cross validation. (D) Impact of data size on model prediction. For data-driven model, the prediction increased with the inclusion of more datasets. For linear model, the prediction decreased when more datasets were used. The shadow areas indicate the 95% confidence interval of model prediction. (E) Validation of the regression tree model by comparing the simulated and experimentally measured gene regulations on Eno2p-tdTomato cassette subject to dCas9-VP64 regulation.

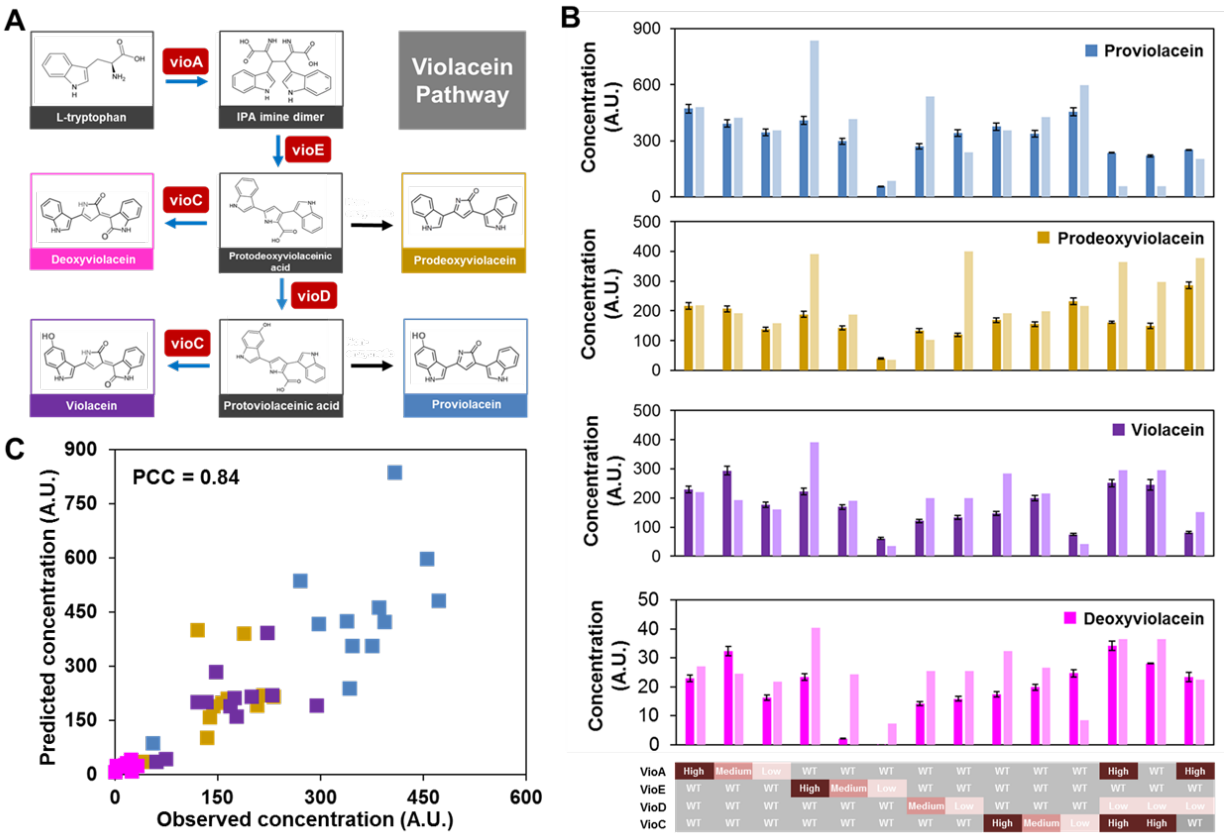

Figure 3. Design dCas9-VP64 to reprogram metabolic fluxes in yeast. (A) Violacein pathway in yeast used in this study to demonstrate the programmable control of metabolic fluxes by using dCas9-VP64. Five enzymatic steps (vioA, vioB, vioC, vioD, and vioE) and two non-enzymatic steps led to four products from the violacein pathway: proviolacein, prodeoxyviolacein, violacein, and deoxyviolacein. (B) A panel of genes subject to regulation of dCas9-VP64 were chosen to control metabolic fluxes to various products from the violacein pathway. WT: wild type gene without any regulation; High: highly up-regulated gene expression by dCas9-VP64; Medium: medium-level up-regulated gene expression by dCas9-VP64; Low: down-regulated gene expression by dCas9-VP64. (C) The compassion between the model-predicted and experimentally measured products from the violacein pathway. The high correlation coefficient

(PCC=0.84) indicated that the data-driven model could accurately predict the effects of artificial Cas9-based regulator on metabolic flux reprogramming.

References

1 Lee, S. K., Chou, H., Ham, T. S., Lee, T. S. & Keasling, J. D. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology* **19**, 556-563, doi:http://dx.doi.org/10.1016/j.copbio.2008.10.014 (2008).

2 Keasling, J. D. & Chou, H. Metabolic engineering delivers next-generation biofuels. *Nature biotechnology* **26**, 298-299, doi:10.1038/nbt0308-298 (2008).

3 Vuoristo, K. S., Mars, A. E., Sanders, J. P., Eggink, G. & Weusthuis, R. A. Metabolic Engineering of TCA Cycle for Production of Chemicals. *Trends Biotechnol* **34**, 191-197, doi:10.1016/j.tibtech.2015.11.002 (2016).

4 Chen, X. *et al.* Metabolic engineering of Escherichia coli: a sustainable industrial platform for bio-based chemical production. *Biotechnology advances* **31**, 1200-1223, doi:10.1016/j.biotechadv.2013.02.009 (2013).

5 Yadav, V. G. & Stephanopoulos, G. Metabolic Engineering: The Ultimate Paradigm for Continuous Pharmaceutical Manufacturing. *ChemSusChem* **7**, 1847-1853, doi:10.1002/cssc.201301219 (2014).

6 Khosla, C. & Keasling, J. D. Metabolic engineering for drug discovery and development. *Nature reviews. Drug discovery* **2**, 1019-1025, doi:10.1038/nrd1256 (2003).

7 Keasling, J. D. Manufacturing Molecules Through Metabolic Engineering. *Science* **330**, 1355-1358, doi:10.1126/science.1193990 (2010).

8 Zadran, S. & Levine, R. D. Perspectives in metabolic engineering: understanding cellular regulation towards the control of metabolic routes. *Appl Biochem Biotechnol* **169**, 55-65, doi:10.1007/s12010-012-9951-x (2013).

9 Zalatan, J. G. *et al.* Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339-350, doi:10.1016/j.cell.2014.11.052 (2015).

10 McNerney, M. P., Watstein, D. M. & Styczynski, M. P. Precision metabolic engineering: The design of responsive, selective, and controllable metabolic systems. *Metabolic engineering* **31**, 123-131, doi:10.1016/j.ymben.2015.06.011 (2015).

11 Broun, P. Transcription factors as tools for metabolic engineering in plants. *Curr Opin Plant Biol* **7**, 202-209, doi:10.1016/j.pbi.2004.01.013 (2004).

12 Kim, J. & Reed, J. L. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol* **4**, 53, doi:10.1186/1752-0509-4-53 (2010).

13 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

14 Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* **31**, 233-239, doi:10.1038/nbt.2508 (2013).

15 Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451, doi:10.1016/j.cell.2013.06.044 (2013).

16 Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature protocols* **8**, 2180-2196, doi:10.1038/nprot.2013.132 (2013).

17 Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183, doi:10.1016/j.cell.2013.02.022 (2013).

18 Farzadfard, F., Perli, S. D. & Lu, T. K. Tunable and multifunctional eukaryotic transcription factors based on CRISPR/Cas. *ACS Synth Biol* **2**, 604-613, doi:10.1021/sb400081r (2013).

19 Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic acids research* **41**, 7429-7437, doi:10.1093/nar/gkt520 (2013).

20 Gao, Y. *et al.* Complex transcriptional modulation with orthogonal and inducible dCas9 regulators. *Nature methods* **13**, 1043-1049, doi:10.1038/nmeth.4042 (2016).

21 Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic acids research* **42**, W401-W407, doi:10.1093/nar/gku410 (2014).

22 Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic acids research* **44**, W272-W276, doi:10.1093/nar/gkw398 (2016).

23 Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach*. (Pearson Education, 2003).

24 Langley, P. The changing science of machine learning. *Machine Learning* **82**, 275-279 (2011).

25 Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of machine learning*. (MIT press, 2012).

26 Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and regression trees*. (CRC press, 1984).

27 Loh, W.-Y. & Shih, Y.-S. Split selection methods for classification trees. *Statistica sinica*, 815-840 (1997).

28 Hoshino, T. Violacein and related tryptophan metabolites produced by Chromobacterium violaceum: biosynthetic mechanism and pathway for construction of violacein core. *Applied microbiology and biotechnology* **91**, 1463-1475, doi:10.1007/s00253-011-3468-z (2011).

29 Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome research* **24**, 132-141, doi:10.1101/gr.162339.113 (2014).

30 Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome biology* **17**, 148, doi:10.1186/s13059-016-1012-2 (2016).

31 Delebecque, C. J., Lindner, A. B., Silver, P. A. & Aldaye, F. A. Organization of intracellular reactions with rationally designed RNA assemblies. *Science* **333**, 470-474, doi:10.1126/science.1206938 (2011).

32 Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLoS Comput Biol* **12**, e1004724, doi:10.1371/journal.pcbi.1004724 (2016).

33 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949, doi:10.1016/j.cell.2014.02.001 (2014).

34 Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997, doi:10.1126/science.1247997 (2014).

35 Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88, doi:10.1126/science.aad5227 (2016).

36 Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867-871, doi:10.1126/science.aad8282 (2016).

37 Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477-1481, doi:10.1126/science.aab1452 (2015).

38 Hittinger, C. T. Saccharomyces diversity and evolution: a budding model genus. *Trends in genetics : TIG* **29**, 309-317, doi:10.1016/j.tig.2013.01.002 (2013).

39 Petzold, C. J., Chan, L. J. G., Nhan, M. & Adams, P. D. Analytics for Metabolic Engineering. *Frontiers in Bioengineering and Biotechnology* **3**, 135, doi:10.3389/fbioe.2015.00135 (2015).

40 Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic acids research* **27**, e16 (2009).

41 Shao, Z., Luo, Y. & Zhao, H. Rapid characterization and engineering of natural product biosynthetic pathways via DNA assembler. *Mol Biosyst* **7**, 1056-1059, doi:10.1039/c0mb00338g (2011).

42 Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic acids research* **41**, 10668-10678, doi:10.1093/nar/gkt809 (2013).