Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Machine learning-assisted medium optimization revealed the discriminated strategies for improved production of the foreign and native metabolites

Honoka Aida [a,1], Keisuke Uchida [a,1], Motoki Nagai [a], Takamasa Hashizume [a], Shunsuke Masuo [a,b], Naoki Takaya [a,b,*], Bei-Wen Ying [a,**]

[a] *School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8572 Ibaraki, Japan*
[b] *Microbiology Research Center for Sustainability, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8572 Ibaraki, Japan*

ARTICLE INFO

Article history:
Received 22 October 2022
Received in revised form 18 April 2023
Accepted 19 April 2023
Available online 20 April 2023

Keywords:
Production
Medium
Synthetic pathway
Machine learning
Bacterial growth
Aromatic compound
Transcriptome

ABSTRACT

The composition of medium components is crucial for achieving the best performance of synthetic construction in genetically engineered cells. Which and how medium components determine the performance, e.g., productivity, remain poorly investigated. To address the questions, a comparative survey with two genetically engineered *Escherichia coli* strains was performed. As a case study, the strains carried the synthetic pathways for producing the aromatic compounds of 4-aminophenylalanine (4APhe) or tyrosine (Tyr), common in the upstream but differentiated in the downstream metabolism. Bacterial growth and compound production were examined in hundreds of medium combinations that comprised 48 pure chemicals. The resultant data sets linking the medium composition to bacterial growth and production were subjected to machine learning for improved production. Intriguingly, the primary medium components determining the production of 4PheA and Tyr were differentiated, which were the initial resource (glucose) of the synthetic pathway and the inducer (IPTG) of the synthetic construction, respectively. Fine-tuning of the primary component significantly increased the yields of 4APhe and Tyr, indicating that a single component could be crucial for the performance of synthetic construction. Transcriptome analysis observed the local and global changes in gene expression for improved production of 4APhe and Tyr, respectively, revealing divergent metabolic strategies for producing the foreign and native metabolites. The study demonstrated that ML-assisted medium optimization could provide a novel point of view on how to make the synthetic construction meet the designed working principle and achieve the expected biological function.

1. Introduction

Designing genetic circuits to construct synthetic metabolic pathways in bacteria is applied for industrial applications and exploring the working principles of living systems [1–4]. In metabolic engineering and synthetic biology, the so-called bottom-up approach using the cellular components as "parts" to rewire the metabolic pathways by genetic reconstruction has been used to produce valuable substances in microorganisms [5,6], e.g., artemisinic acid production [7]. Various tools have been developed to achieve native metabolites or foreign substrates [8], such as the utilization of genome-scale metabolic models (GEMs) [9] and phenotype prediction by machine learning (ML) [10]. Incorporating the synthetic gene circuits into the native regulatory mechanisms has been challenged in building industrially valuable synthetic pathways and creating the synthetic metabolism of novel functions [11–13].

In addition to the genetic design, the culture medium is crucial for achieving the best performance of the synthetic construction and modified metabolic pathway [14]. Medium optimization has usually been performed to improve the productivity of the endpoint compound (e.g., metabolite, protein) [15–17]. The approaches for medium optimization have been intensively developed [18]. The methods based on personal experience and knowledge are popular

[*] Corresponding author at: School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8572 Ibaraki, Japan.
[**] Corresponding author.
E-mail addresses: takaya.naoki.ge@u.tsukuba.ac.jp (N. Takaya), ying.beiwen.gf@u.tsukuba.ac.jp (B.-W. Ying).

[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2023.04.020

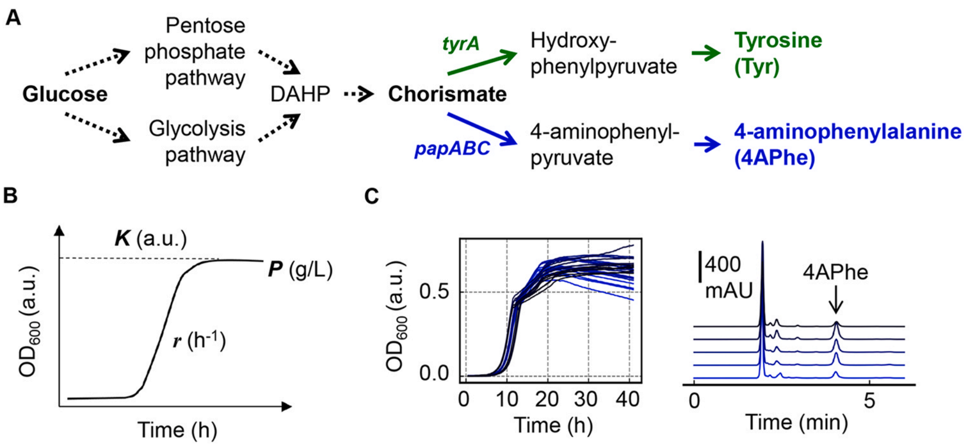

Fig. 1. An overview of the experimental and computational analyses. A. Flow chart of the synthetic pathways for producing the aromatic compounds. Blue and green indicate the metabolic flows of producing 4APhe and Tyr in two different *E. coli* strains, respectively. B. The three parameters used in the study. *K*, *r*, and *P* represent the maximal density, growth rate, and production yield. The units of the three parameters are indicated. C. Examples of the analytical results of bacterial growth and production. The Left and right panels show the growth curves and chromatography peaks in various medium combinations, respectively.

for quick laboratory decisions. Those based on the statistical and computational approaches, e.g., response surface methodology (RSM) [19–21], have been used for improved development. Since metabolism is a highly complex network, ML has been employed to achieve more significant and efficient optimization [22–25]. Progress in well-established high-throughput technologies and automation primarily benefited the acquisition of large datasets [26,27], which are essential for ML. The previous studies demonstrated that introducing ML to medium optimization successfully accelerated bacterial growth [28–30] and increased productivity [31,32].

Despite the successes in genetic design and medium optimization for metabolic engineering, how the optimized medium improved the function of the synthetic pathways (i.e., the production of metabolites) remained largely unknown. For instance, it's unclear which medium component primarily decided the production of metabolites via the synthetic pathways. As the changes in medium combinations would trigger transcriptome reorganization [33,34], how the optimized medium contributed to the transcriptional changes of the genes that participated in the synthetic pathways for improved production was unclear. The present study discovered the primary medium components for improved production of synthetic constructions and investigated the changes in gene expression triggered by the fine-tuned medium composition. As a pilot study, two genetically engineered *Escherichia coli* strains producing two different aromatic compounds, i.e., the native and foreign metabolites, via the synthetic pathways of the common upstream and differentiated downstream, were subjected to medium optimization. High-throughput assays were performed to link the medium combinations to bacterial growth and production. ML was applied to the resultant datasets to find the medium components improving productivity. Transcriptome analysis was conducted to discover the metabolic strategy for the increased yields of synthetic constructions.

2. Results

2.1. Synthetic pathways for producing aromatic compounds

To test the availability of ML to medium optimization for synthetic construction, the genetically engineered *E. coli* strains producing 4-aminophenylalanine (4APhe) [35] or tyrosine (Tyr) [36] were used as a case study. The strains were previously developed and carried the synthetic pathways that were common from glucose to chorismate and differentiated from chorismate to the endpoint metabolites (Fig. 1A). As the wild-type *E. coli* could biosynthesize Tyr but not 4APhe, the compounds of Tyr and 4APhe were designated as the native and foreign metabolites, respectively. Thousands of growth curves were acquired by high-throughput growth assay, and two representative parameters of the growth rate (*r*) and maximal population density (*K*) were calculated in accordance (Fig. 1B). High-performance liquid chromatography (HPLC) was performed to evaluate the production (*P*) of the two compounds at the stationary phase (Fig. 1C). Note that the experimental and analytical approaches used here were all well-established in our previous studies as described in the Materials and Methods.

2.2. Linking medium combinations to bacterial growth and production

A total of 48 pure compounds, including buffers, sugars, nitrogens, sulfurs, amino acids, metal compounds, vitamins, antibiotics, and inducers, were used to prepare various medium combinations for examining bacterial growth and production. Note that 44 out of 48 compounds were adopted from the previous study using wild-type *E. coli* [28]. Because of the synthetic construction, three antibiotics (chloramphenicol, ampicillin, and streptomycin) and an inducer (IPTG) were added in the present study. 192 and 378 medium combinations were tested for the strains producing 4APhe and Tyr, respectively (Table S1, Table S2). As the pure compounds were ionized in solution, the medium combinations finally consisted of 44 chemical components (Fig. S1, Fig. S2). Repeated assay resulted in thousands of growth curves, and the mean values were used for ML. The final datasets of *r*, *K*, and *P* were 189, 192, and 192 values for 4APhe and 377, 378, and 378 for Tyr. A considerable variation in growth and production was acquired in both strains; however, their distributions were different (Fig. 2A). The distributions of *r*, *K*, and *P* all presented monomodal shapes for the strain producing 4APhe (Fig. 2A, upper panels). In comparison, the distribution of *r* was bimodal, although those of *K* and *P* remained monomodal for the strain producing Tyr (Fig. 2A, bottom panels). The dissimilarity indicated that the two synthetic constructions differed in the linkages between growth and production.

The relationship between growth and production was additionally investigated by correlation analysis. The production (*P*) of 4APhe was positively correlated to the growth rate (*r*) and population density (*K*), whereas no correlation was detected between *r* and *K* (Fig. 2B, upper panels). Statistically significant correlations were commonly observed between any pair of *r*, *K*, and *P* in the strain producing Tyr (Fig. 2B, bottom panels), indicating that growth and

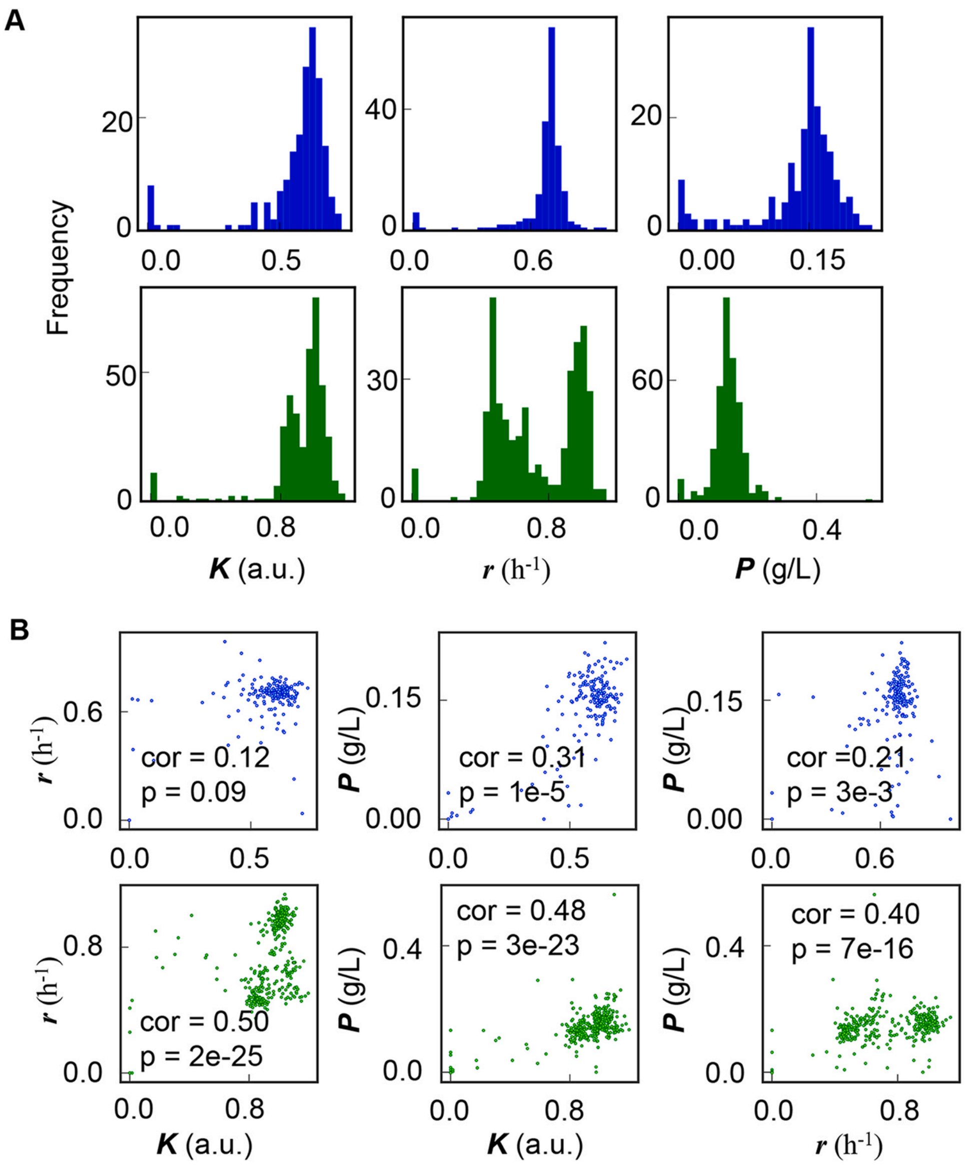

Fig. 2. Bacterial growth and production in hundreds of medium combinations. A. Profiling of the growth and productivity. Histograms of the maximal density, growth rate, and production yield of the *E. coli* strains grown in various medium combinations are shown from left to right, respectively. The numbers of the tested medium combinations for the *E. coli* strains producing 4APhe and Tyr are 192 and 378, respectively. B. Relationships between growth and production. *K*, *r*, and *P* represent the maximal density, growth rate, and production yield. Scatter plots of any pair of *K*, *r*, and *P* are shown. Spearman's correlation coefficients and the p-values are indicated. Blue and green represent the results of the *E. coli* strains holding the synthetic constructions for producing 4APhe and Tyr, respectively.

production have a trend to change in relation. The associations among *r*, *K*, and *P* were highly significant in producing the native metabolite, Tyr, and became weaker in producing the foreign one, 4APhe.

2.3. Predicting the primary medium components for production

Before predicting the medium components determinative for bacterial growth and productivity, six representative ML models, i.e., gradient-boosted decision tree (GBDT), classification and regression trees (CART), linear regression, support vector machine (SVM), random forest, and extreme gradient boosting (XGB), were compared. The results demonstrated that GBDT presented higher prediction accuracy than others (Fig. 3A), which agreed with our previous study [28]. Accordingly, GBDT was employed for the following analysis. The priority of the medium components (i.e., feature importance) in deciding *r*, *K*, and *P* was evaluated. The results showed that glucose was of high priority in deciding the growth and production (Fig. 3B), which was reasonable as glucose was the initial resource for both aromatic compounds (Fig. 1A). The primary components deciding *K* and *r* were glucose and magnesium in the strain producing 4APhe (Fig. 3B, upper panels); however, they were both glucose in the strain producing Tyr (Fig. 3B, bottom panels). The findings were consistent with the relationship between *r* and *K*, which was no correlation in the strain producing 4APhe and positively correlated in the strain producing Tyr (Fig. 2B, left panels).

Intriguingly, the medium components most influencing the production of 4APhe and Tyr were different, i.e., glucose and IPTG, respectively (Fig. 3B, right panels). According to the synthetic pathways, glucose was the common initial resource, and IPTG was the common inducer for producing both compounds. Different strategies might be adopted in the two strains to fine-tune the synthetic pathways for production. That is, the metabolic resource-oriented and the transcriptional regulation-dominated strategies in producing 4APhe and Tyr, respectively. In addition, approximately equivalent numbers of the components determined the growth (*r*, *K*)

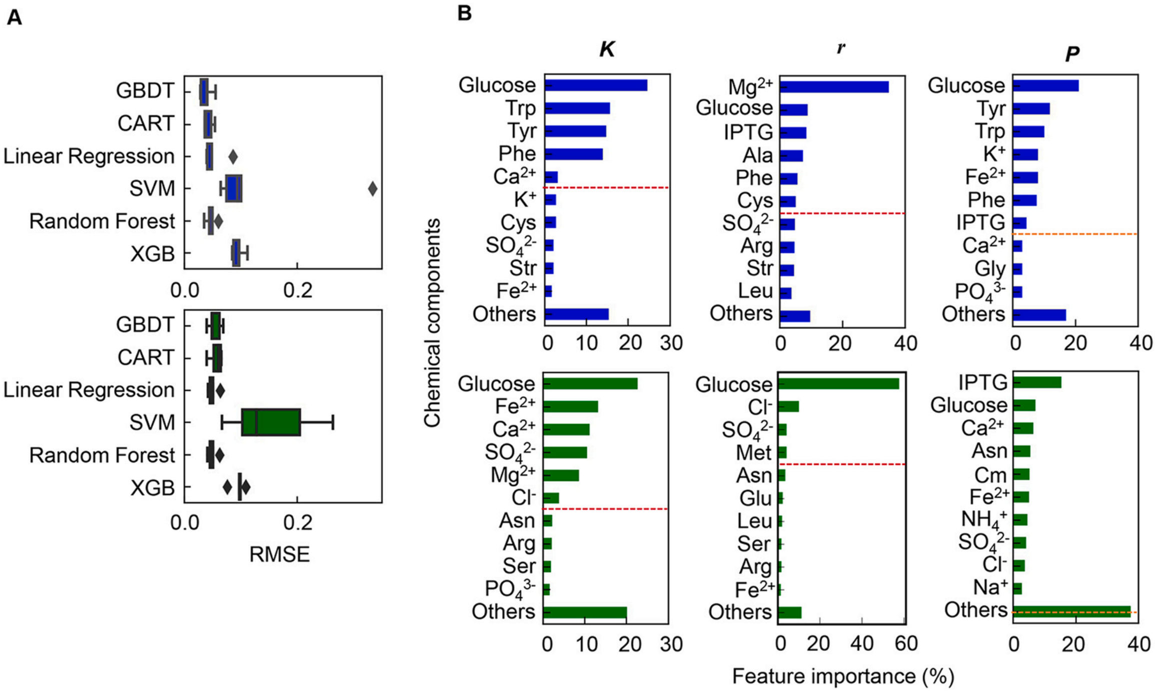

Fig. 3. Contribution of the medium components to growth and production. A. Prediction accuracy of six ML models. Root mean squared errors (RMSEs) of five independent tests are shown for each ML. Blue and green indicate the productivity 4APhe and Tyr, respectively. B. The contribution of each medium component to the maximal density, growth rate, and productivity was predicted by the GBDT model. The top ten primary components affecting the three parameters are displayed, and the rest components are summarized as "Others". *K*, *r*, and *P* represent the maximal density, growth rate, and production yield. Blue and green represent the results of the *E. coli* strains holding the synthetic pathways for producing 4APhe and Tyr, respectively. The broken lines in orange indicate the border of the cumulated feature importance of ~70 %, as summarized in Table S3.

of the two strains (Fig. 3B, broken lines, ~70 % of feature importance); however, the numbers were significantly different in deciding the productivities of 4APhe and Tyr, i.e., 7 and 14 components, respectively (Fig. 3B, Table S3). It revealed that more components participated in producing the native metabolite than the foreign one.

2.4. Fine-tuning the concentration of the medium components for improved production

Classification and regression tree (CART) was subsequently applied to estimate the range of chemical concentration, as described previously [30]. The components associated with their concentrations were predicted, and those of high priority in contributing to compound production (*P*) were differentiated in the two strains (Table 1). Glucose and IPTG were the primary components in deciding the production of 4APhe and Tyr, respectively. Trp and Phe were predicted as the decision-making components for producing 4APhe were reasonable, as the synthetic pathways disturbed the native metabolism of tryptophan and phenylalanine biosynthesis [35]. In comparison, the two amino acids insignificantly contributed to the production of Tyr, although their biosynthesis was interrupted by the synthetic pathways. The differentiation in the primary components for substrate production in the two strains was highly consistent with that predicted with GBDT, despite a few dissimilarities between the two ML models.

Table 1
Chemical concentrations and productivity predicted by CART. The high-priority components contributing to productivity (*P*) and the predicted concentrations of the components for improved productivity are indicated. Note that Cu^{2+} and PABA were of low priority, predicted by GBDT.

	4APhe	Tyr
Concentration (mM)	Trp > 0.06 **Glucose > 6.3** Cu^{2+} > 1.0×10^{-5} Phe > 0.02	**0.01 < IPTG < 0.03** PABA > 0.001
P (g/L)	0.15	0.56

Medium optimization by fine-tuning the concentrations of glucose and IPTG was performed to demonstrate whether the predicted chemical concentrations improved the production. The results showed that the yields of both compounds were changed in response to the concentration gradients of the two primary components (Fig. 4 A). The chemical concentrations presenting high yields were precisely within the predicted ranges. The concentrations of glucose and IPTG for the best production of 4APhe and Tyr were finally decided as 40 and 0.025 mM, respectively. Additional experiments verified the yields of 4APhe and Tyr were 0.30 and 0.65 g/L, which were higher than the prediction of 0.15 and 0.56 g/L, respectively (Table 1).

Furthermore, whether the fine-tuning of the primary components improved the production in the culture of a larger volume was examined. A scaleup to 5 ML of culture was conducted to compare the production of 4APhe and Tyr in three different media, i.e., the fine-tuned, pre-fine-tuned, and M9-LB media. M9-LB was a commonly used rich medium optimized in the previous studies that constructed the two *E. coli* strains [35,36]. The highest yields of both 4APhe and Tyr were acquired in the fine-tuned media (Fig. 4B, left panels), comparable to those cultured in 200 μL (Fig. 4A). Intriguingly, the increased yields linked to the lower growth rates and maximal population densities (Fig. 4B, right two panels). It strongly suggested that the trade-off between growth and production should be considered in synthetic constructions.

2.5. Changes in gene expression caused by the fine-tuned medium composition

Transcriptome analysis was performed to investigate whether and how a single fine-tuned medium component triggered the changes in gene expression. Differential expression genes (DEGs)

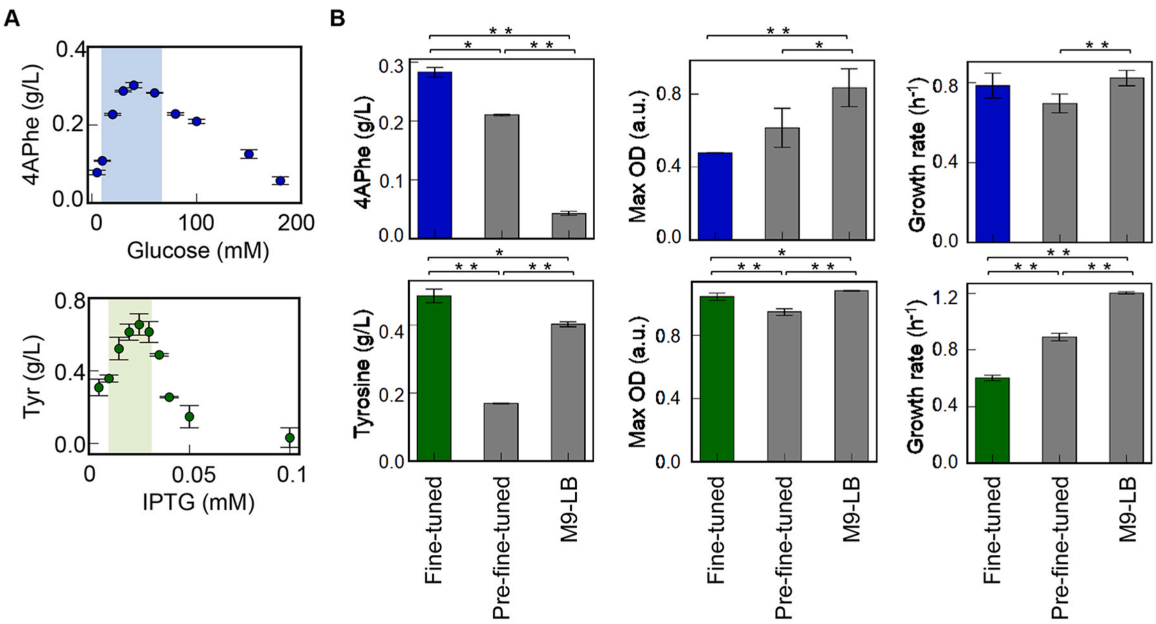

Fig. 4. Changes in production in response to medium variation. A. Testing the primary components for production. The upper and bottom panels indicated the production of 4APhe and Tyr, respectively. Ten medium combinations of identical concentrations of 47 components but varying concentrations of glucose or IPTG were tested. The shadows represent the concentration gradients of the primary components (glucose or IPTG) predicted by the CART model. Standard errors of experimental replications are indicated. B. Testing the representative media. The media of fine-tuned and pre-fine-tuned differed in the concentration of the primary component, i.e., glucose for 4APhe and IPTG for Tyr. M9-LB represents the mixture of the M9 and LB media used previously. Asterisks represent the statistical significance (*, $p < 0.05$; **, $p < 0.01$). Standard errors of experimental replications are indicated.

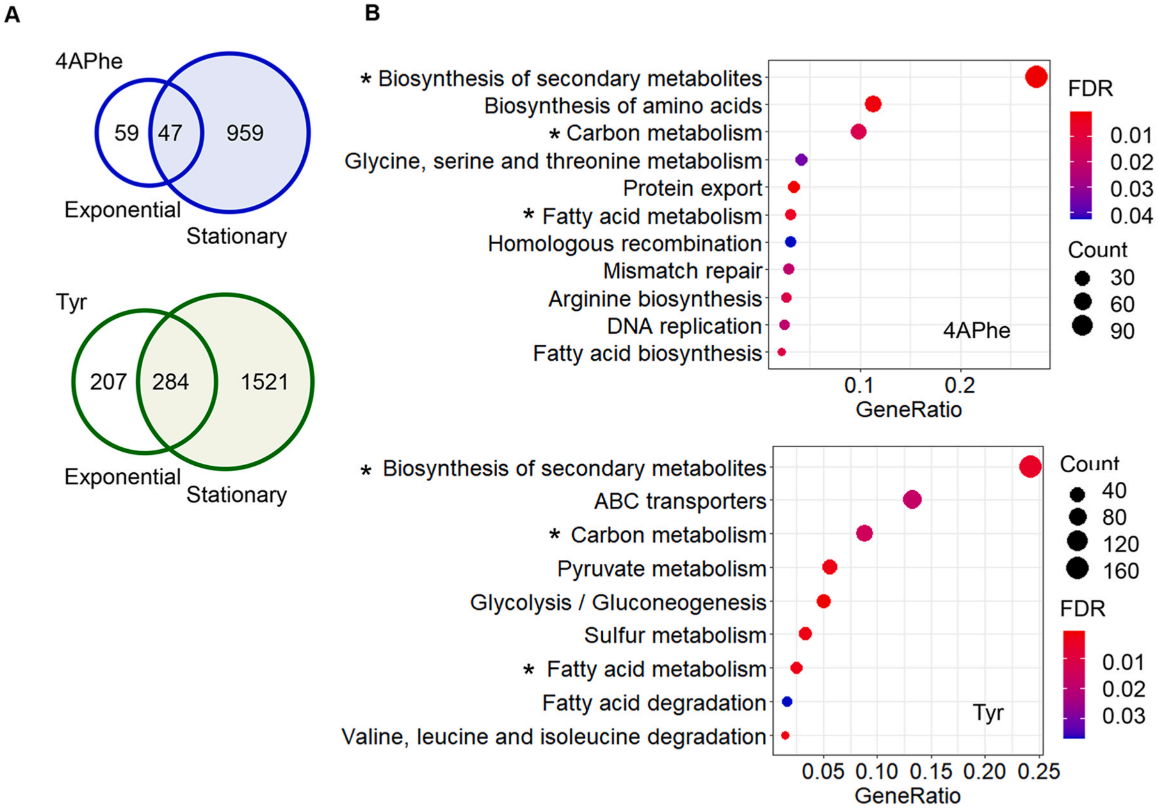

Fig. 5. Differentiated gene expression in response to medium optimization. A. Numbers of DEGs in response to medium optimization. Open and shadowed circles stand for the exponential and stationary phases, respectively. Blue and green represent the results of the *E. coli* strains holding the synthetic pathways for producing 4APhe and Tyr, respectively. B. Enriched metabolic pathways in DEGs. The KEGG pathways significantly enriched in the DEGs are indicated (FDR < 0.05). The upper and bottom panels indicate the strains producing 4APhe and Tyr, respectively. Gene ratio is the ratio of DEGs to all genes annotated in the pathway. Color gradation and the size of circles indicate the statistical significance represented by FDR and the number of DEGs annotated in the pathway.

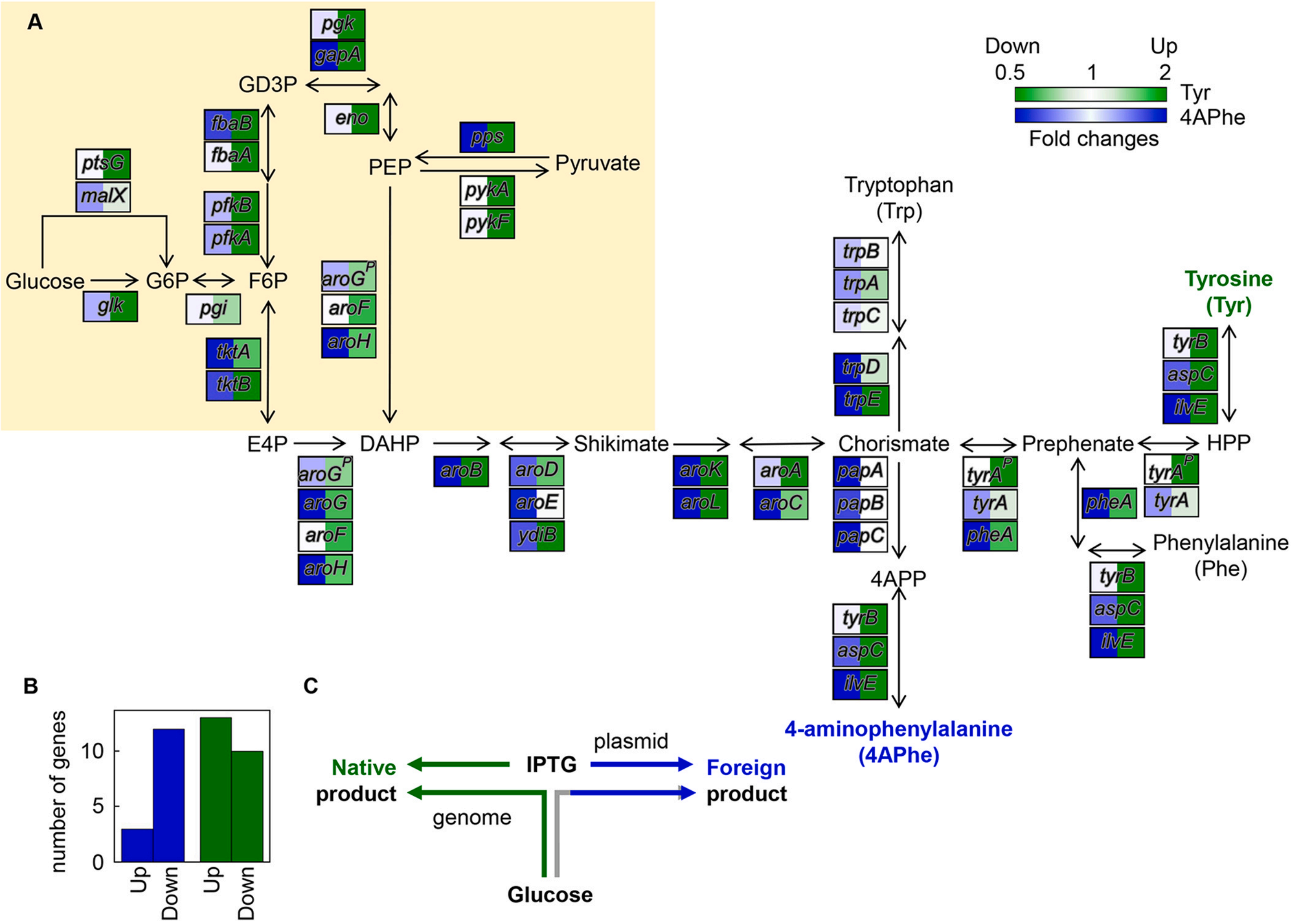

Fig. 6. Transcriptional changes in the synthetic pathways. A. Transcriptional changes of the genes participated in the metabolic pathways for production. The genes and chemical compounds participating in the metabolic pathways are shown. Color gradation in blue and green represent the fold changes of gene expression caused by the medium optimization. Shadow in light yellow indicates the upstream glycolysis and pentose phosphate pathways in common for producing 4APhe and Tyr. B. Number of genes of transcriptional changes. The genes participating in the metabolic pathway, which showed fold changes (⟦$\log_2$FC⟧ > 1), were counted. Upstream and downstream indicate the genes shadowed in light yellow (A) and the others, respectively. Blue and green represent the strains producing 4APhe and Tyr, respectively. C. Schematic drawing of the working principle of synthetic pathways.

mediated by medium optimization were identified more in the stationary phase than in the exponentially growing phase (Fig. 5A). Only a few DEGs overlapped between the exponential and the stationary phases. Significant changes in gene expression commonly occurred in the phase for production but not growth, independent of the variation in the fine-tuned components (i.e., glucose and IPTG) and final products (i.e., 4APhe and Tyr). It indicated that ML-assisted prediction was practical for finding the critical component for efficient medium optimization, and tuning a single medium component was sufficient to improve the performance of the synthetic construction.

Functional enrichment analysis of DEGs showed that a total of 11 and 9 KEGG pathways significantly (FDR < 0.05) fluctuated in the strains producing 4APhe and Tyr, respectively (Fig. 5B). However, the numbers of DEGs varied between the two strains, i.e., 1006 and 1805 DEGs. Three native metabolic pathways (biosynthesis of secondary metabolites, carbon metabolism, and fatty acid metabolism) were commonly enriched in response to the medium optimization (Fig. 5B, asterisks), possibly due to the universal upstream of the synthetic pathways for the production of 4APhe and Tyr.

In addition, the improved production of 4APhe was mainly attributed to the transcriptional regulation of downstream pathways close to the endpoint metabolite (Fig. 6A, blue). In comparison, the improved production of Tyr was accompanied by transcriptional changes across entire pathways from the initial resource to the endpoint metabolite (Fig. 6A, green). Despite the universal upstream metabolism, the expression changes of the genes that participated in the synthetic pathways occurred at the local and global scale for 4APhe and Tyr, respectively (Fig. 6B). It indicated that the transcriptome reorganization for improved production was differentiated between the foreign and native metabolites. Since the genomic backgrounds to harbor the synthetic constructions were not identical, the differentiated mechanisms might not be wholly due to the synthetic pathways. Nevertheless, whether the fine-tuned component was a resource initiating the metabolism or an inducer regulating the gene expression must be crucial for metabolic strategy (Fig. 6C).

3. Discussion

Combining the data-driven approach and the transcriptome analysis allowed us to find the different strategies of metabolic optimization between foreign and native metabolites. The highest priority chemicals contributing to the production of 4 APhe and Tyr were glucose and IPTG, respectively. The genes for aromatic amino acids biosynthesis and transcriptional feedback repression (i.e., *tyrA*, *trpE*, *pheA*, and *tyrR*) were mutated or deleted in the strains producing 4APhe [35,37]. The feedback repression was released so that the limiting factor for 4APhe production became the initial resource, glucose. On the other hand, as the feedback mechanism remained regular in the strain producing Tyr, the IPTG-induced expression level of the enzyme (i.e., *tyrA*) might play an essential role.

Fine-tuning of a single medium component caused the transcriptome reorganization of the overall pathways for Tyr production, indicating that balancing metabolism was essential for producing the native metabolite. Since the native proteins and metabolites participate in multiple cellular processes, global activation of all these processes might cause metabolic imbalance and be lethal, which was supported by the fact that the overexpression of native proteins caused severe inhibition of *E. coli* growth [38]. It might be why the inducer IPTG was predicted to be the primary component for producing Tyr. A balanced metabolism mediated by gene expression at an appropriate induction level must have been crucial for producing native compounds (metabolite). In addition, only the local but not global reorganization of the transcriptome occurred in producing the foreign compound, 4APhe. The local metabolic modification might be sufficient for improved production because of the fewer interactions of the foreign metabolite to the native pathways. It was supported by the previous study that found the metabolic imbalance occurring in the overexpression of native but not orthologous dihydrofolate reductase (DHFR) [39].

In the present study, adding the inducer at the beginning of the bacterial culture seemed to work well, as it caused much more DEGs in the stationary phase than in the exponential phase. The appropriate amount of inducer at the initial growth phase could benefit the transcriptome for better production in the stationary phase, which was consistent with the study on high penicillin productivity without growth burden in *E. coli* [40]. Nevertheless, an alternative study reported that adding IPTG in the early exponential growth phase could cause a significant burden on growth and protein production in *E. coli* [41]. The timing of IPTG induction might depend on the target products/metabolites and synthetic constructions. The optimal IPTG concentration for Tyr production (Fig. 4A) was ~ 10-fold lower than those generally used for recombinant protein production [42,43]. The improved production of the synthetic construction was not simply due to the induced expression of the critical enzyme but the transcriptional reorganization of the whole metabolic pathway (Fig. 6). To fully understand the mechanism of improved productivity, further analysis of the glycolysis and pentose pathways and the metabolic reactions related to ATP and reductants are required.

Whether the primary component changed in the culture medium of a larger volume was tested. The production of 4APhe was highly responsive to glucose in a 5 ML culture (Fig. S4) within the concentration gradient as in the 200 µL culture (Fig. 4A). It demonstrated that glucose played a primary role in producing 4APhe in a large-volume culture. In addition, the fine-tuned medium was highly competitive for producing 4APhe compared to the commercially available rich media (Fig. S5). Despite the highest yield in the fine-tuned medium, the growth rate and maximal population density were lower than those in the commercial media (Fig. S5). The trade-off between production and growth demonstrated that the fine-tuned medium was directed explicitly toward production but not growth.

The tested medium combinations were limited in the present study; however, the high-throughput assay combined with ML-assisted medium optimization was practical. High nutritional richness was assumed to benefit bacterial growth and production, which was not true in the present case (Fig. S3). Identifying the primary components for improved growth or production was crucial for efficient medium optimization. So far, finding an ideal medium for the synthetic construction to reach its best performance remains challenging because both the medium and the metabolism are highly complex. Whether the optimized media were the best remained unclear without direct comparison to others, despite that they performed better than the traditional rich medium (Fig. 4B). The ML-assisted optimization allowed 48 components to be adjusted simultaneously, which was robust compared to other methods. As a case study, the present study provided a pilot trial of medium optimization by exploring components that improve bacterial production with synthetic pathways and offered valuable hints for the multidisciplinary approaches for the field. For instance, developing the culture media and optimizing the culture conditions for producing multidomain or modified components [44,45] require considering numerous factors to find the balanced combination that satisfies highly complicated biochemical reactions. Improving the ML models through the continuous biotechnological application will promote optimizing complex media for complex compounds.

In addition, the ML models were constructed with the concentration gradients across a wide range changing on a logarithmic scale. Despite of high performance demonstrated initially (Fig. 3A), they might be unavailable to provide accurate predictions in response to slight changes in chemical concentration. Testing the experimental data of a narrow concentration gradient within an order and changing on a linear scale (Fig. 4A) showed poor prediction accuracy (Fig. S6). As the experimental results used for the model training (Fig. 2A) were significantly lower than those for the test (Fig. 4A), the decision tree algorithms were somehow weak at the prediction out of the range of the training data. These issues could be addressed by increasing the training data with a broad range of changes at a high density of intervals, which might be costed in time and labor. An initial ML-assisted optimization followed by manual fine-tuning was supposed to be practical in the laboratory and industry. Note that the ML-assisted medium optimization did not address the scaleup issue commonly occurring in industrial applications, which was independent of the methodologies and a general limitation for medium optimization.

In summary, the present study showed that the fine-tuned media led to differentiated transcriptional changes for better production of foreign and native compounds. It strongly suggested that the metabolic strategies for producing foreign and native were different. An increased initial resource (e.g., glucose) could improve the endpoint productivity (e.g., 4APhe), probably because of the weak interaction between the foreign product and the native metabolism. A balanced induction of metabolic pathways was essential for improved productivity of native metabolites (e.g., Tyr), which might be caused by the tight regulation affecting the overall metabolic balance. These insights were valuable for using the bacterial cell as a factory in synthetic biology and significantly benefited from the ML-assisted high-throughput approaches used here.

4. Materials and methods

4.1 E.coli strains

Two genetically reconstructed *E. coli* strains that produced two different aromatic compounds from chorismate via the shikimic acid pathway were used. The *E. coli* strain NST37(DE3)/Δ*pheLA* harboring pET-pfpapA/pCDF-pfpapBC/pACYC-aroG4 was previously constructed to produce 4-aminophenylalanine (4APhe) production [35]. The expression of *papABC* mediated the production of 4APhe. The *E. coli* strain BL21(DE3) harboring pET-tyrA/pACYC-aroG^{fbr} was constructed to produce tyrosine (Tyr), as described previously [36]. The low-copy plasmids were used for the synthetic construction, and both products were induced by isopropyl β-D-1-thiogalactopyranoside (IPTG) (Fig. 1A).

4.2. Natural media

The media of Instant LB, APS, Plusgrow II, and Lennox were obtained commercially (BD Biosciences), and M9LB was prepared in the lab. The M9LB medium comprises 24 g/L Na_2HPO_4, 12 g/L KH_2PO_4, 1 g/L NH_4Cl, 0.5 g/L NaCl, 0.05 g/L thiamine HCl, 0.5 g/L $MgSO_4$-7 H_2O, 0.015 g/L $CaCl_2$-H_2O, 5 g/L Yeast extract, 10 g/L Tryptone and 2 ML/L Hunter's trace elements, which contain 22 g/L $ZnSO_4$-H_2O, 11 g/L H_3BO_3, 5 g/L $MnCl_2$-4 H_2O, 5 g/L $FeSO_4$-7 H_2O, 1.6 g/L $CoCl_2$-6 H_2O and $CuSO_4$/5 H_2O. The chemical compounds are all commercially available (Wako). 0.2 mM IPTG was added to these rich media to induce the production of 4APhe.

4.3. Preparation of medium combinations

A total of 48 pure compounds were used, of which 44 were previously determined [28], and four were newly added. The four additives were the inducer (i.e., IPTG) and antibiotics (i.e., ampicillin, chloramphenicol, and streptomycin). The pure compounds were all purchased commercially (Wako). The concentration gradients of these compounds were determined as the same as in the previous study [28]. In brief, the minimal and maximal concentrations were zero and 10–100 folds of the concentrations present in the commonly used media, respectively. The chemical concentrations were varied on a logarithmic scale. The stock solutions of these 48 compounds were prepared and stored in small aliquots (100–1000 µL) at − 30 °C, as described previously. The stock solutions were used only once, and the remainder was discarded to avoid deterioration of compound quality due to repeated thawing and freezing. The medium combinations were prepared by mixing these stock solutions before the growth assay. Finally, 192 and 378 medium combinations were tested in the growth assays for the *E. coli* strains that produce 4APhe and Tyr, respectively (Table S1).

4.4. Bacterial growth assay

The cell stocks of exponentially growing *E. coli* cells ($OD_{600} \approx 0.01 - 0.1$) were prepared before the growth assay for repeated use, as described previously [46]. The cell stocks (cell culture aliquots) were used only once, and the remaining culture was discarded. The high-throughput growth assays were performed using the 96-well microplates (Costar) and the plate reader (Epoch2, BioTek), as described in the previous studies [28]. Every four to six wells (200 µL per well) in different positions were tested for each medium combination. The temporal changes of the *E. coli* cells growing at 37 °C were recorded by reading absorbance at 600 nm for 30–48 h in 30-min intervals. In addition, the growth assay in a test tube was performed with 5 ML of the culture at 37 °C and 200 rpm using a bioshaker (Taitec). The cell culture was temporally sampled by taking 0.1–1.0 ML of culture to measure OD600 (Beckman DU730), as previously described in detail [47].

4.5. Growth data processing and calculation

The temporal OD_{600} reads were exported from the plate reader and processed with Python. The growth parameters *r* and *K* were evaluated according to previous reports [30,48] using a previously developed Python program [30]. In brief, *r* was defined as the mean of three continuous logarithmic slopes of every two neighboring OD_{600} values within the exponential growth phase using "gradient" in the "NumPy" library. *K* was calculated as the mean of three continuous OD_{600} values, including the maximum, determined using "argmax" in the "NumPy" library. The mean of biological replicates (four to six micro-well cultures) was used for *r* and *K*.

4.6. High-performance liquid chromatography

The amounts of 4APhe and Tyr produced by the *E. coli* cells grown in various media were evaluated by high-performance liquid chromatography (HPLC). The mixture of the biological replicates (four to six micro-well cultures) was subjected to HPLC to achieve the average productivity of each medium combination. The mixed cultures were centrifuged at 16,000 rpm for 5 min (Model 3700, Kubota), and the supernatants were collected and subjected to HPLC (1260 Infinity, Agilent Technologies). The flow rate of HPLC was set at 0.8 ML/min. A HiQ sil C18HS-3 column (Siri Instrument) was equipped to detect 4APhe. The initial mobile phase was 20 mM KH_2PO_4 (pH 7.0) and pure methanol (Sigma-Aldrich) in the ratio of 98–2 and maintained for 7 min. The methanol concentration was gradually increased to 50 % in 5 min and then maintained for 5 min. Subsequently, the ratio of methanol was gradually reduced to 2 % for 2 min, and the flow was kept for 4 min. In addition, a TSKgel ODS-100 V column (Tosoh) was used to detect Tyr. The initial mobile phase was 10 mM ammonium formate (pH 7.0) and pure acetonitrile (Merck Millipore) in the ratio of 95–5 and maintained for 8 min. The acetonitrile concentration was increased to 50 % in 6 min and then maintained for 2 min. Subsequently, the ratio of acetonitrile was gradually reduced to 5 % in 4 min, and the flow was kept for 5 min *p*-Amino-L-phenylalanine (Sigma-Aldrich) and L-tyrosine (Wako) were used as the standards of 4APhe and Tyr, respectively. The amounts (productivity, *P*) of 4APhe and Tyr were calculated according to the peak areas at the corresponding retention times.

4.7. Machine learning

Machine learning (ML) was performed using Python with some modifications from previous studies [28,30]. Six ML models, i.e., gradient-boosted decision tree (GBDT), classification and regression trees (CART), linear regression, support vector machine (SVM), random forest, and extreme gradient boosting (XGB), were tested to compare the prediction accuracy of productivity (*P*). GBDT and CART were finally applied to the datasets that linked growth parameters (*r* and *K*) and productivity (*P*) to the medium combinations. The chemical concentrations of the medium combinations were transformed into logarithmic values in the datasets. "GradientBoostingRegressor" in the "ensemble" module, "DecisionTreeRegressor" in the "tree" module, "LinearRegression" in the "linear_model" module, 'SVR' in the 'svm' module, and 'RandomForestRegressor' in the 'ensemble' module were used for GBDT, CART, linear regression, SVM, and random forest, respectively. These modules were all in the "scikit-learn" library. "XGBResressor" in the "xgboost" library was used for XGB. Fivefold nested cross-validation was performed to evaluate the ML models, and root mean squared error (RMSE) was used to evaluate accuracy. As previously described [28], the datasets were split for training and evaluation. Outer cross-validation (outer cv) and inner cross-validation (inner cv) were applied to evaluate the ML accuracy and determine the hyperparameters. A grid search was used for the hyperparameter search as follows. In GBDT and XGB, "n_estimators" was set to 300. In addition, "learning_rate" and "max_depth" were searched from 0.005 to 0.5 in increments of 0.005 and from 1 to 4 in increments of 1, respectively. In CART, "criterion" and "max_depth" were set to "mse" and 5, respectively. In SVM, 'kernel' was set to 'rbf'. Additionally, 'C', 'gamma', and 'epsilon' were searched from 2^{-5} to 2^{10}, 2^{-20} to 2^{10}, and 2^{-10} to 2^{0}, respectively, in increments of 2^{2}. In random forest, 'random_state' and 'n_estimators' were set to 0 and 300, respectively, and 'max_depth' was searched among 2, 3, and 4. All other hyperparameters were used as default. The "feature_importance_" values were calculated by fivefold cross-validation. The mean value of five repeated calculations was used as the GBDT predicted output.

4.8. RNA purification and RNA sequencing

The *E. coli* cells were cultured in 5 ML of the fresh media and rotated at 200 rpm at 37 °C. Independent cultures as biological replicates of transcriptomes were performed. The cell cultures of both exponential and stationary growth phases were collected, as described previously [49,50]. The total RNAs were purified using RNeasy Mini Kit (QIAGEN) and RNase-Free DNase Set (QIAGEN) according to the product instructions. The eluted RNAs were suspended with RNase-free water and subsequently subjected to RNAseq. The rRNAs were removed using Ribo-Zero Plus rRNA Depletion Kit (Illumina), and mRNA preparation was performed with Ultra Directional RNA Library Prep Kit for Illumina (NEBNext). The paired-end sequencing (150 bp × 2) was performed with Novaseq6000 (Illumina) by Chemical Dojin Co. Ltd.

4.9. Transcriptome data processing and analysis

The computational and statistical analyses were all performed using R [51]. The reference genome sequences of *E. coli* W3110 and BL21(DE3) were obtained from the GenBank of the accession numbers ASM1024v1 and CP001509.3, respectively. The trimmed reads were mapped to the reference genome sequences with Bowtie2 [52]. The transcriptome datasets were deposited in the DNA Data Bank of Japan (DDBJ) under the accession number DRA013628. The differential expressed genes (DEGs) were determined using DESeq2 package [53], with a criterion of $FDR < 0.05$ [54], where the read counts were directly used, as previously described [55]. The enrichment analysis was performed using the globally normalized datasets, as previously described [49,56]. KEGG pathways [57,58] were enriched using "enrichKEGG" in clusterProfiler [59], where "pAdjustMethod" was set to "fdr," and "organism" was set to "ecj" and "ebe" for the *E. coli* strains that produced 4APhe and Tyr, respectively.

CRediT authorship contribution statement

KU and MN performed the experiments; HA, KU, MN, TH and BWY analysed the data; HA, TH, SM, NT and BWY developed experimental and analytical tools; NT and BWY lead the project; BWY conceived the research; HA and BWY wrote the manuscript; and all authors approved the final paper.

Competing interest

The authors (BWY and HA) have competing interests. The medium combinations were related to a patent under the control number 2021-171528 (Japan).

Acknowledgments

This work was partially supported by the JSPS KAKENHI Grant-in-Aid numbers 19H03215 and 21K19815 to BWY, and the JSPS KAKENHI grant "Post-Koch Ecology" (19H05679 and 19H05687) to NT.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.04.020.

References

[1] Otero-Muras I, Carbonell P. Automated engineering of synthetic metabolic pathways for efficient biomanufacturing. Metab Eng 2021;63:61–80.

[2] Long B, Fischer B, Zeng Y, Amerigian Z, Li Q, Bryant H, et al. Machine learning-informed and synthetic biology-enabled semi-continuous algal cultivation to unleash renewable fuel productivity. Nat Commun 2022;13:541.

[3] Fong SS. Computational approaches to metabolic engineering utilizing systems biology and synthetic biology. Comput Struct Biotechnol J 2014;11:28–34.

[4] Jouhten P. Metabolic modelling in the development of cell factories by synthetic biology. Comput Struct Biotechnol J 2012;3:e201210009.

[5] Stephanopoulos G. Synthetic biology and metabolic engineering. ACS Synth Biol 2012;1:514–25.

[6] Cameron DE, Bashor CJ, Collins JJ. A brief history of synthetic biology. Nat Rev Microbiol 2014;12:381–90.

[7] Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 2006;440:940–3.

[8] Keasling JD. Synthetic biology and the development of tools for metabolic engineering. Metab Eng 2012;14:189–95.

[9] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol 2019;20:121.

[10] Carbonell P, Radivojevic T, García Martín H. Opportunities at the intersection of synthetic biology, machine learning, and automation. ACS Synth Biol 2019;8:1474–7.

[11] Satowa D, Fujiwara R, Uchio S, Nakano M, Otomo C, Hirata Y, et al. Metabolic engineering of E. coli for improving mevalonate production to promote NADPH regeneration and enhance acetyl-CoA supply. Biotechnol Bioeng 2020;117:2153–64.

[12] Huccetogullari D, Luo ZW, Lee SY. Metabolic engineering of microorganisms for production of aromatic compounds. Micro Cell Fact 2019;18:41.

[13] Larroude M, Celinska E, Back A, Thomas S, Nicaud JM, Ledesma-Amaro R. A synthetic biology approach to transform Yarrowia lipolytica into a competitive biotechnological producer of β-carotene. Biotechnol Bioeng 2018;115:464–72.

[14] Overmann J, Abt B, Sikorski J. Present and future of culturing bacteria. Annu Rev Microbiol 2017;71:711–30.

[15] Abuhena M, Al-Rashid J, Azim MF, Khan MNM, Kabir MG, Barman NC, et al. Optimization of industrial (3000 L) production of Bacillus subtilis CW-S and its novel application for minituber and industrial-grade potato cultivation. Sci Rep 2022;12:11153.

[16] Krause M, Neubauer A, Neubauer P. The fed-batch principle for the molecular biology lab: controlled nutrient diets in ready-made media improve production of recombinant proteins in Escherichia coli. Micro Cell Fact 2016;15:110.

[17] Choi GH, Lee NK, Paik HD. Optimization of medium composition for biomass production of Lactobacillus plantarum 200655 using response surface methodology. J Microbiol Biotechnol 2021;31:717–25.

[18] Singh V, Haque S, Niwas R, Srivastava A, Pasupuleti M, Tripathi CK. Strategies for fermentation medium optimization: an in-depth review. Front Microbiol 2016;7:2087.

[19] Aguirre AM, Bassi A. Investigation of biomass concentration, lipid production, and cellulose content in Chlorella vulgaris cultures using response surface methodology. Biotechnol Bioeng 2013;110:2114–22.

[20] Bezerra MA, Santelli RE, Oliveira EP, Villar LS, Escaleira LA. Response surface methodology (RSM) as a tool for optimization in analytical chemistry. Talanta 2008;76:965–77.

[21] Latha S, Sivaranjani G, Dhanasekaran D. Response surface methodology: a non-conventional statistical tool to maximize the throughput of Streptomyces species biomass and their bioactive metabolites. Crit Rev Microbiol 2017;43:567–82.

[22] Packiam KAR, Ooi CW, Li F, Mei S, Tey BT, Ong HF, et al. PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in Escherichia coli. Comput Struct Biotechnol J 2022;20:2909–20.

[23] Lawson CE, Martí JM, Radivojevic T, Jonnalagadda SVR, Gentz R, Hillson NJ, et al. Machine learning for metabolic engineering: a review. Metab Eng 2021;63:34–60.

[24] Kim GB, Kim WJ, Kim HU, Lee SY. Machine learning applications in systems metabolic engineering. Curr Opin Biotechnol 2020;64:1–9.

[25] Cuperlovic-Culf M. Machine learning methods for analysis of metabolic data and metabolic pathway modeling. Metabolites 2018;8.

[26] Gilpin W, Huang Y, Forger DB. Learning dynamics from large biological data sets: machine learning meets systems biology. Curr Opin Syst Biol 2020;22:1–7.

[27] Suthers PF, Foster CJ, Sarkar D, Wang L, Maranas CD. Recent advances in constraint and machine learning-based metabolic modeling by leveraging stoichiometric balances, thermodynamic feasibility and kinetic law formalisms. Metab Eng 2021;63:13–33.

[28] Aida H, Hashizume T, Ashino K, Ying BW. Machine learning-assisted discovery of growth decision elements by relating bacterial population dynamics to environmental diversity. Elife 2022;11.

[29] Hiura S, Koseki S, Koyama K. Prediction of population behavior of Listeria monocytogenes in food using machine learning and a microbial growth and survival database. Sci Rep 2021;11:10613.

[30] Ashino K, Sugano K, Amagasa T, Ying BW. Predicting the decision making chemicals used for bacterial growth. Sci Rep 2019;9:7251.

[31] Kumar P, Adamczyk PA, Zhang X, Andrade RB, Romero PA, Ramanathan P, et al. Active and machine learning-based approaches to rapidly enhance microbial chemical production. Metab Eng 2021;67:216–26.

[32] Zheng Z-Y, Guo X-N, Zhu K-X, Peng W, Zhou H-M. Artificial neural network – genetic algorithm to optimize wheat germ fermentation condition: Application

to the production of two anti-tumor benzoquinones. Food Chem 2017;227:264–70.

[33] Feugeas JP, Tourret J, Launay A, Bouvet O, Hoede C, Denamur E, et al. Links between transcription, environmental adaptation and gene variability in Escherichia coli: correlations between gene expression and gene variability reflect growth efficiencies. Mol Biol Evol 2016.

[34] Blair JMA, Richmond GE, Bailey AM, Ivens A, Piddock LJV. Choice of bacterial growth medium alters the transcriptome and phenotype of salmonella enterica serovar typhimurium. PLOS ONE 2013;8:e63912.

[35] Masuo S, Zhou S, Kaneko T, Takaya N. Bacterial fermentation platform for producing artificial aromatic amines. Sci Rep 2016;6:25764.

[36] Masuo S, Saga C, Usui K, Sasakura Y, Kawasaki Y, Takaya N. Glucose-derived raspberry ketone produced via engineered Escherichia coli metabolism. Front Bioeng Biotechnol 2022;10:843843.

[37] Tribe DE. Novel microorganism and method. United States: Austgen Biojet International Pty Ltd; 1987.

[38] Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H, et al. Complete set of ORF clones of Escherichia coli ASKA library (A Complete S et of E. coli K -12 ORF A rchive): unique resources for biological research. DNA Res 2005;12:291–9.

[39] Bhattacharyya S, Bershtein S, Yan J, Argun T, Gilson AI, Trauger SA, et al. Transient protein-protein interactions perturb E. coli metabolome and cause gene dosage toxicity. eLife 2016;5:e20309.

[40] Ramírez OT, Zamora R, Espinosa G, Merino E, Bolívar F, Quintero R. Kinetic study of penicillin acylase production by recombinant E. coli in batch cultures. Process Biochem 1994;29:197–206.

[41] Malakar P, Venkatesh KV. Effect of substrate and IPTG concentrations on the burden to growth of Escherichia coli on glycerol due to the expression of Lac proteins. Appl Microbiol Biotechnol 2012;93:2543–9.

[42] Lipničanová S, Legerská B, Chmelová D, Ondrejovič M, Miertuš S. Optimization of an Inclusion Body-Based Production of the Influenza Virus Neuraminidase in Escherichia coli. Biomolecules 2022;12:331.

[43] Einsfeldt K, Severo Júnior JB, Corrêa Argondizzo AP, Medeiros MA, Alves TLM, Almeida RV, et al. Cloning and expression of protease ClpP from Streptococcus pneumoniae in Escherichia coli: Study of the influence of kanamycin and IPTG concentration on cell growth, recombinant protein production and plasmid stability. Vaccine 2011;29:7136–43.

[44] Soares EV. Perspective on the biotechnological production of bacterial siderophores and their use. Appl Microbiol Biotechnol 2022;106:3985–4004.

[45] Garrigues L, Do TD, Bideaux C, Guillouet SE, Meynial-Salles I. Insights into Clostridium tetani: from genome to bioreactors. Biotechnol Adv 2022;54:107781.

[46] Kurokawa M, Ying BW. Precise, high-throughput analysis of bacterial growth. J Vis Exp 2017.

[47] Tsuchiya K, Cao YY, Kurokawa M, Ashino K, Yomo T, Ying BW. A decay effect of the growth rate associated with genome reduction in Escherichia coli. BMC Microbiol 2018;18:101.

[48] Liu L, Kurokawa M, Nagai M, Seno S, Ying BW. Correlated chromosomal periodicities according to the growth rate and gene expression. Sci Rep 2020;10:15531.

[49] Ying BW, Matsumoto Y, Kitahara K, Suzuki S, Ono N, Furusawa C, et al. Bacterial transcriptome reorganization in thermal adaptive evolution. BMC Genom 2015;16:802.

[50] Ying BW, Seno S, Kaneko F, Matsuda H, Yomo T. Multilevel comparative analysis of the contributions of genome reduction and heat shock to the Escherichia coli transcriptome. BMC Genom 2013;14:25.

[51] Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.

[52] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.

[53] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

[54] Storey JD. A direct approach to false discovery rates. J R Stat Soc: Ser B (Stat Methodol) 2002;64:479–98.

[55] Matsui Y, Nagai M, Ying BW. Growth rate-associated transcriptome reorganization in response to genomic, environmental, and evolutionary interruptions. Front Microbiol 2023;14:1145673.

[56] Ying BW, Yama K. Gene expression order attributed to genome reduction and the steady cellular state in Escherichia coli. Front Microbiol 2018;9:2255.

[57] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457–62.

[58] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61.

[59] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.