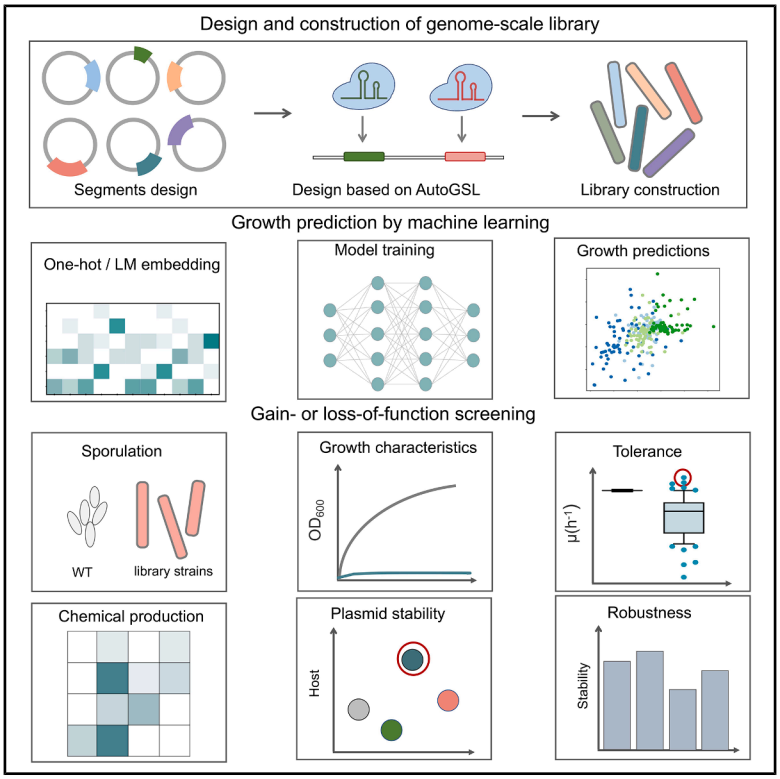
iScience

Article

Chromosome-segment scanning for gain- or loss-of-function screening (CHASING)

Graphical abstract



Highlights

- We developed a genome-scale library and applied it to rapidly screen phenotypes
- The gene-knockout induced phenotype would not be shadowed by the unrelated gene deletion
- The specific gene behind an observed phenotype could be quickly identified by CHASING

Authors

Yan Xia, Lichao Sun, Zeyu Liang, ..., Pengyu Dong, Yi-Xin Huo, Shuyuan Guo

Correspondence

huoyixin@bit.edu.cn (Y.-X.H.), guosy@bit.edu.cn (S.G.)

In brief

Genomic library; Sequence analysis

Xia et al., 2025, iScience *28*, 112484
May 16, 2025
https://doi.org/10.1016/j.isci.2025.112484

CellPress

iScience

CellPress
OPEN ACCESS

Article

Chromosome-segment scanning for gain- or loss-of-function screening (CHASING)

Yan Xia,[1] Lichao Sun,[1,2] Zeyu Liang,[1] Zhongrao Han,[1] Jing Li,[1] Pengyu Dong,[1] Yi-Xin Huo,[1,2,3,*] and Shuyuan Guo[1,4,*]
[1]Department of Gastroenterology, Aerospace Center Hospital, College of Life Science, Beijing Institute of Technology, No. 5 South Zhongguancun Street, Haidian District, Beijing 100081, China
[2]Tangshan Research Institute, Beijing Institute of Technology, No. 57, South Jianshe Road, Lubei District, Tangshan, Hebei 063000, China
[3]Center for Future Foods, Muyuan Laboratory, Zhengzhou, Henan 450016, China
[4]Lead contact
*Correspondence: huoyixin@bit.edu.cn (Y.-X.H.), guosy@bit.edu.cn (S.G.)
https://doi.org/10.1016/j.isci.2025.112484

SUMMARY

Identifying beneficial or functional mutations for a specific species is a task always relies on the labor-intensive construction of a library containing thousands of single-gene knockout or interference strains. Here, we systematically demonstrated that the task could be done by constructing a genome-scale library, designed by AutoGSL, containing a limited number of large fragment deletion strains. The loss-of- and gain-of-function phenotypes could be efficiently screened out by our chromosome-segment scanning, and the specific gene corresponding to an observed phenotype could be quickly identified and visualized by our computer-aided gene-annotation web tool. Additionally, a Clusters of Orthologous Gene Transformer learned representations were transferred to predict growth phenotypes of the genome-scale library under varying conditions. We further utilized our chromosome segment scanning for gain- or loss-of-function screening (CHASING) strategy to obtain acetoin- and lycopene-overproducing hosts. Our work highlighted the significance of CHASING in functional genomics investigation, robust chassis engineering, and chemical overproduction.

INTRODUCTION

Native bacteria maintain homeostasis through the regulation of the expression of the functional proteins at transcriptional and translational levels. The deactivation of a gene could lead to the loss or destroy of a function, while the deactivation of a regulator could lead to function restore via de-repression or de-regulation of a previously depressed enzyme or pathway. Dissecting the genotype-phenotype map, which implies the relationship between genotype variations and phenotype changes,[1] provides valuable insights into the engineering of robust chassis and is essential for industrial chemical production.

Developing high-throughput approaches is important for understanding the sequence-function as well as gene-network architecture.[2] Constructing randomly generated mutant libraries has been widely employed using transposon-induced mutation strategies.[3] However, the transposon-mediated mutation resulted in insertion bias toward genes with long coding regions.[4,5] With the advance of the CRISPR technique, the pooled CRISPR interference (CRISPRi) technique has been applied for screening by designing the genome-scale sgRNA libraries in microorganisms such as *Escherichia coli*,[5,6] *Synechocystis* sp PCC 6803[7] and *Eubacterium limosum*,[8] despite the limitations such as the variation of the binding affinity of sgRNAs and the fluctuating repression level during sampling.[5,9,10]

An alternative high-throughput approach for mapping phenotypes is the genome-arrayed single-gene mutant libraries containing thousands of strains. By knocking out the non-essential genes individually, collections of single-gene-deletion strains have been reported for elaborating genome architecture and stability in *E. coli*,[11,12] *Bacillus subtilis*,[13] and *Saccharomyces cerevisiae*.[14] Single-gene-knockdown strains have also been obtained using the CRISPRi technique, revealing a systematic essential-gene network of *B. subtilis*.[15] Although exploring these libraries avoids the unknown targeting effect, the library construction requires enormous labor work, and the operating processes for screening and storage are challenging.[16,17]

To provide an alternative simple strategy for large-scale genotype-phenotype mapping, we sought to establish an arrayed chromosome-segment-deletion (CSD) library and group the genes within segments to a specific cluster of function. First, we developed an AutoGSL gene editing tool to design genome-scale libraries. We yielded a genome-scale library consisting of 70 CSD strains, each consisting of a functional grouping of non-essential genes for *B. subtilis* by AutoGSL and applied them to rapidly screen the loss-of- and gain-of-function phenotypes. We employed COG Transformer encoding to train machine learning models for predicting genome-scale library growth phenotypes, demonstrating its ability to establish a functional mapping from the genome. By taking advantage of the well-established database of Clusters of Orthologous Gene (COG), we showed that the CSD library could have a 55.7% of probability to link a single gene directly to an observed phenotype in *B. subtilis* 168. We also evaluated the applicability of

the CSD library in metabolic engineering applications. Specifically, we successfully screened out two lycopene-overproducers reaching 54.1% and 18.2% increase in yields, and five sporulation-defective mutants reaching as high as 20.8% increase in acetoin titers. Our research underscored that the simultaneous disruption of several genes exhibited negligible effects on the genes responsible for associated phenotypic traits, highlighting the significance and generality of the CHASING strategy in diverse applications.

RESULTS

Constructing a genome-scale deletion library in *B. subtilis* by AutoGSL

Here, we sought to investigate whether a gene-knockout induced phenotype could still be screened out in the presence of other unrelated gene deletions by constructing an arrayed chromosome-segment-deletion (CSD) library. There are 257 essential genes and 4018 non-essential genes in the genome of *B. subtilis* 168 (Figure S1**)**. If each essential gene were set as a dividing point, the whole genome could be divided into 257 segments. Taking into account the 109 cases where no non-essential gene exists between two essential genes, we finally divided the whole genome, by essential genes, into 148 large segments, each containing a group of non-essential genes. Since the 96-well plate has the greatest convenience and potential for large-scale genotype-phenotype mapping, constructing a library of 70–80 strains is a reasonable choice considering the positive control, negative control, and necessary blank.

The construction of a CDS library requires iterative design of editing sequences, including sgRNAs and homologous arms, to achieve precise knockout of target sequences. Manual design of knockout sequences for the library is not only low in throughput and prone to errors, but also risks overlooking critical quality control steps. To address this, we developed the AutoGSL tool, which automates the design of sgRNAs, homologous arms, and primers required for experiments (Figure 1A). Users simply input the names of the first and last genes of the target fragment, and AutoGSL will automatically generate the gene editing design. For example, if we want to knock out the genes rlbA to gyrB, AutoGSL identifies three suitable sgRNAs based on PAM sequences, GC content, and editing efficiency (https://chasingdesign.cloudmol.org/v1/design_md/?start_gene=rlbA&end_gene=gyrB). Additionally, the tool designs the Left HR and Right HR primers for PCR based on GC content, T_m values, and homologous arm length. The application of AutoGSL significantly reduces the design workload for constructing a CDS library.

Taking advantage of the ten *Bacillus* large-fragment knockout strains previously available in our lab,[18] we designed a 75-CSD library for *B. subtilis* to provide a proof-of-concept demonstration for large-scale screening. In this study, 65 chromosome-segment deletions were performed individually (Figure 1B). For each segment, two sgRNAs targeting both ends and a repair template were designed using AutoGSL (Figure 1C). Among the segments, 48 were deleted with success rates over 60%, including 31 with 100% success (Figure 1D). The segment sizes ranged from 3.7 to 201 kb (Figure 1E), showcasing the high efficiency of our large-fragment deletion method. Despite numerous attempts with diverse sgRNAs, five segments (Δ4, Δ12, Δ26, Δ40, and Δ62) ranging from 5.6 to 22.7 kb could not be deleted, suggesting that certain non-essential genes or the combination of non-essential genes could not be knocked out. These fragments cannot be knocked out, maybe because they involve core pathways essential for maintaining life, such as central carbon metabolism and amino acid metabolism. Taken together, we have yielded a genome-wide library consisting of 70 CSD strains, including 60 strains that were obtained in this work and ten strains that we reported earlier.

To explore the genomic features of this large-scale CSD library, we evaluated the effects of the deleted genes on the metabolic network. We established the interaction network for the proteins translated from the deleted genes. We performed statistical analysis on the degree of the genes that were not knocked out, the knocked-out genes, and the degree between the knocked-out and non-knocked-out genes. The results demonstrated that non-knockout genes had a higher degree within the overall protein network of the strain, whereas knocked-out genes exhibited more intimate interactions with the remainder of the network (t-test p-value = 3.38×10^{-43}) (Figure 1F). A highly crosstalk has been observed in the whole protein map. The simultaneous deletions of several nodes are unlikely to cause a catastrophic effect on the network due to the function crosstalk among remaining genes in the absence of the deleted genes, reflecting the high robustness of our deletion library. This robustness was further supported by the high interconnectivity of proteins transcribed from genes within each segment (Figure S2). To visualize the protein functional distribution, we applied t-SNE to project proteins using machine-learned representations derived from the ProtT5, a widely used protein language model (pLM). Representations derived from pLMs were widely used in protein functional prediction, including GO terms[19] and EC numbers.[20] Proteins that share similar functions are likely to be positioned closer in the projected 2D space (Figure 1G). Proteins encoded by the deleted genes (green points) and the remaining genes (gray points) were uniformly distributed, indicating no significant bias in the design of the deletion segments.

Screening the genome-wide library for loss- and gain-of-function phenotype

To demonstrate the utility of this library in large-scale screening, we screened out the sporulation- and growth-deficient strains for loss-of-function phenotypes, and the various stress-resistant strains for gain-of-function phenotypes. First, sporulation-deficient strains Δ5, Δ15, Δ24, and Δ34 were identified by impaired sporulation in LB medium (Figure S3). Next, the survival of 70 strains in LB medium was assessed, with most showing growth rates similar to the wild-type strain, indicating no significant impact on growth in nutrient-rich conditions (Figures 2A and S4A). The specific growth rates of deletion strains ranged from 0.107 to 0.148 h^{-1}, with the wild-type at 0.142 h^{-1}. Strains Δ15, Δ24, Δ34, and Δ39 showed slower growth after 12 h, resulting in lower 24-h growth rates (0.121, 0.118, 0.113, and 0.107 h^{-1}, respectively). Interestingly, Δ15, Δ24, and Δ34 also exhibited sporulation deficiencies, suggesting a link between spore formation and late stationary-phase growth in *B. subtilis*.

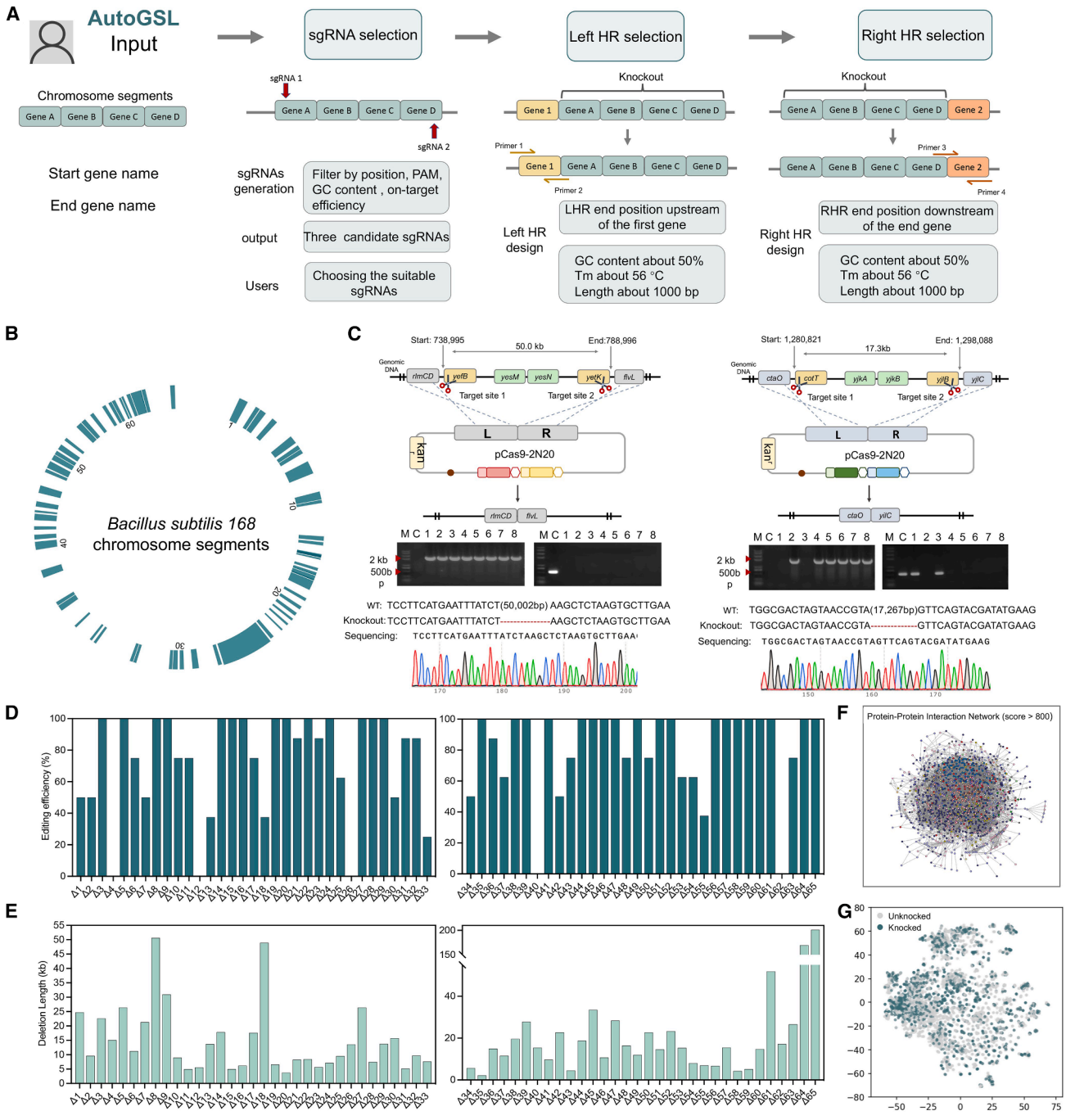


Figure 1. Design and construction of a genome-wide chromosome-segment-deletion library in *B. subtilis* 168 by AutoGSL

(A) Schematic of AutoGSL workflow.

(B) 65 chromosome segments divided by essential genes.

(C) Design and validation of 50.0 kb- and 17.3 kb-chromosome-segment-deletion.

(D) The editing efficiency of 65 chromosome-segment deletions.

(E) The size of the 65 deleted segments constructed in this study.

(F) The interaction network of the proteins translated from the deleted genes. All available data on protein-protein interactions in *B. subtilis* were retrieved from the STRING database. Data of protein-protein interactions with a combined score greater than 800 were considered reliable. Each point represents a protein, and the lines connecting the points depict the interaction.

(G) The functional distribution of proteins in a two-dimensional space. The high-dimensional data were embedded onto a two-dimensional space by the t-SNE projections. The distance of inter-points reflects the function similarity. Gray and green points represent the proteins encoded by the undeleted and deleted genes, respectively.

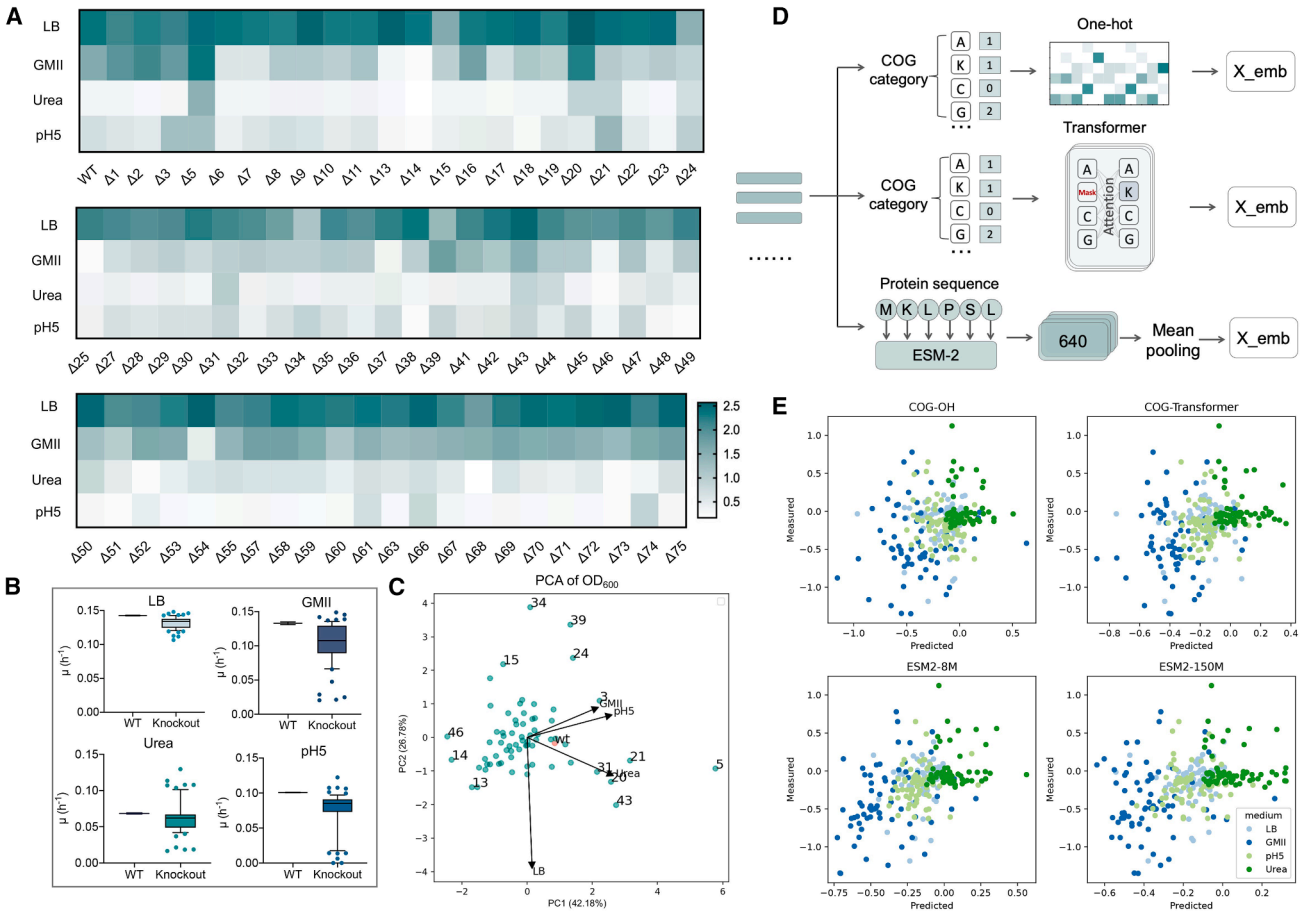


Figure 2. Cellular fitness of the genome-wide deletion library and prediction by machine learning

(A) Heatmap representation of morphological patterns of chromosome-segment-deletion strains in LB, GMII, high urea, and pH 5 medium. Data are means ± SEM.

(B) The specific growth rates of the deletion strains in LB, GMII, urea, and pH 5 medium. The best and the worst growth rates were also displayed in the scatterplot.

(C) Principal component analysis (PCA) of the growth fitness under different conditions.

(D) Three encoding methods for machine learning models.

(E) Growth prediction performance of COG-OH, COG-Transformer, and ESM2-8M/150M in four different media conditions.

Genome-wide identification of conditionally essential genes is crucial for understanding metabolic pathways and network regulation. To achieve this, we screened the library in GMII medium to identify genes essential for growth in nutrient-poor conditions. Most deletion strains exhibited normal growth, with an OD_{600} of around 1.5 at 24 h, similar to the wild-type (Figures 2A and S4B). However, four strains, Δ13, Δ14, Δ25, and Δ46, showed minimal growth with specific growth rates below 0.02 h^{-1}, indicating that the deleted chromosome segments likely contain genes or pathways critical for growth in nutrient-poor conditions (Figure 2B).

B. subtilis is an efficient chassis for exogenous protein expression and is commonly used for urease secretion to hydrolyze urea into NH_3. The ability to utilize high urea concentrations is crucial for this process, but elevated urea can alter osmotic pressure, damaging cell membranes and proteins, and reducing *B. subtilis* survival. While normal growth was seen at 40 and 50 g/L urea, growth was significantly inhibited at 60 g/L (Figure S5A). To enhance urea tolerance, we tested growth under 60 g/L urea stress using the CSD library. Strains Δ5, Δ20, Δ21, Δ31, and Δ43 showed improved growth rates of 0.130, 0.108, 0.107, 0.109, and 0.104 h^{-1}, respectively (Figures 2B and S6A), much higher than the wild-type strain (0.068 h^{-1}), indicating their potential for efficient urea utilization.

During industrial fermentation, organic acids accumulate, causing intracellular acid shock that cannot be fully mitigated by extracellular pH adjustment. Thus, screening for strains tolerant to acid accumulation-induced low pH is crucial. To validate the genome-scale deletion library, we assessed the survival of library strains in GMII medium under acidic conditions. Wild-type *B. subtilis* was first tested at pH 3, 4, and 5. No growth occurred at pH 3 or 4, indicating complete inhibition (Figure S5B), while partial growth was observed at pH 5. Therefore, pH 5 was chosen to screen the CSD library for low pH tolerance. Strains Δ5 and Δ21 demonstrated increased fitness at pH 5, reaching OD_{600} values of 1.20 and

1.69, respectively, compared to wild-type *B. subtilis* (Figures 2A and S6B).

To visualize chassis fitness variance, we conducted principal component analysis (PCA) by plotting the growth data in different conditions along the two most descriptive principal components (Figure 2C). The variable points were scattered near the origin, forming four distinct groups with varied growth patterns. By projecting data points onto the first principal component, we identified three groups: group one, containing strain Δ5 with the highest fitness under GMII, pH 5, and urea-rich conditions; group two, including strains Δ21, Δ43, Δ20, Δ31, and Δ3, which outperformed the wild-type strain (red dot in PCA) but were inferior to Δ5; and group three, consisting of strains with the poorest growth in those conditions. The second principal component highlighted group four, consisting of strains with poor growth in nutrient-rich conditions such as LB medium.

Genome-wide library growth prediction by machine learning

To better predict the growth phenotypes of *B. subtilis* mutant strains, we trained models based on existing experimental data. We employed three protein-based feature extraction methods for genes within large genomic segments: 1) encoding each protein using COG one-hot encoding; 2) encoding the proteins in the entire segment using COG Transformer; and 3) encoding each protein using the language models ESM2-8M/150M and applying average pooling to obtain the feature vector for each segment (Figure 2D). For the COG one-hot encoding, we downloaded 1,053 *B subtilis* genomes from NCBI and re-annotated the COG classifications of each protein using eggnog-emapper, representing the genome as a sequence of COG category strings. For the COG Transformer encoding, we used a Transformer encoder with BERT-style pretraining, training on the COG string sequences of the 1,053 genomes using masked language modeling. The feature vectors generated from this encoding were used as inputs to an Extra Trees model, with ΔOD_{600} measured under four different culture media conditions as the labels. We performed 5-fold cross-validation within each media condition and aggregated the predictions from all test sets to calculate the correlation with experimental data. The results showed that machine learning models trained with different encoding methods could, to some extent, predict bacterial growth under various environmental conditions. The Spearman correlation coefficients for the COG one-hot encoding, COG Transformer, ESM2-8M, and ESM2-150M models were 0.193, 0.326, 0.420, and 0.337, respectively (Figure 2E). Compared to directly using COG categories, the COG Transformer was able to capture a coarser representation of protein function, with performance approaching that of fine-grained protein sequence encodings. This advantage stems from the pretraining process and the attention mechanism of the Transformer, which captures functional semantic information within genomic segments. While the ESM2 protein language model, which computes on a per-amino-acid basis, further improved predictive performance, the COG Transformer encoding was more computationally efficient. With further expansion of the pretraining scope, COG Transformer could become a powerful tool for large-scale genomic feature extraction.

Bioinformatic analysis of the genome-wide deletion library

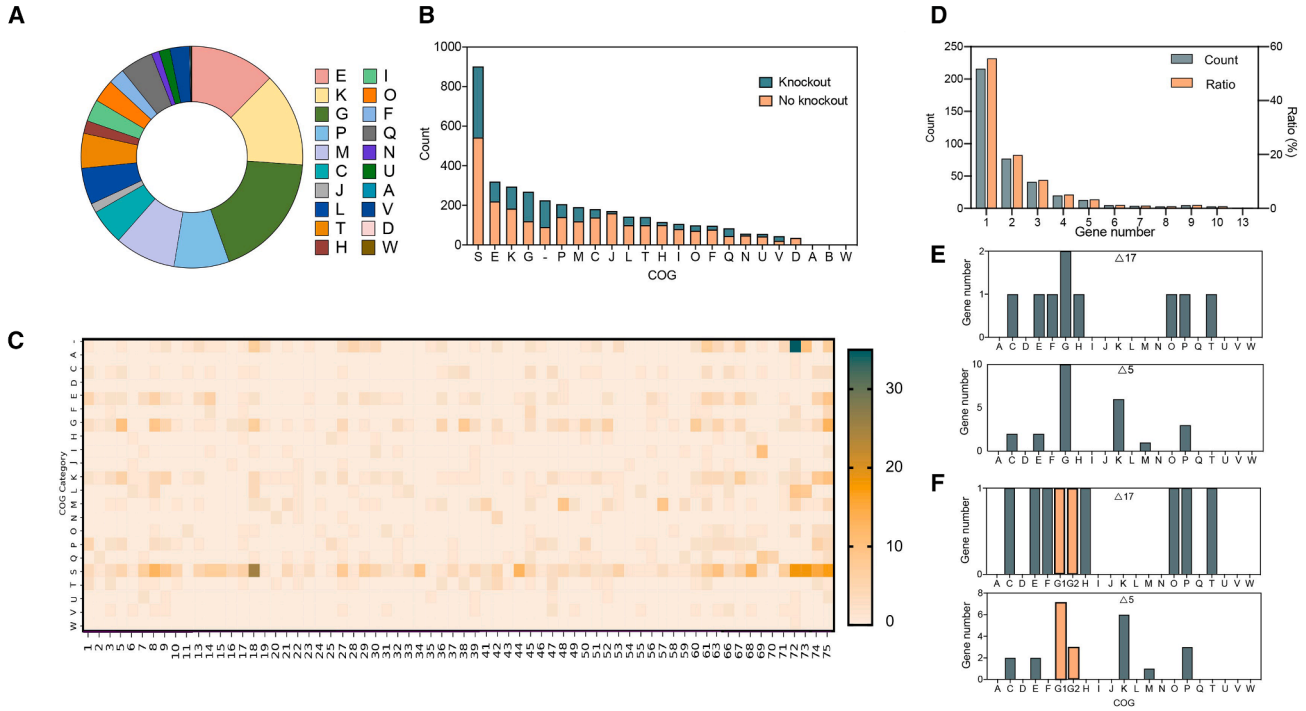
The ability of the COG Transformer encoding method to predict the growth of mutant strains suggests that the COG functional classification has biological significance and encapsulates a wealth of information. The database of COGs includes 26 COG categories such as amino acid metabolism and transport, transcription, translation, cell wall/membrane/envelop biogenesis, and signal transduction.[21] If a segment does not contain an essential gene, it can be deleted, and the loss of non-essential genes may result in specific phenotypes. If a desirable phenotype for metabolic engineering emerges, it is likely due to the loss of genes from a COG linked to that phenotype, allowing rapid identification of the relevant gene.

Based on this concept, we performed bioinformatic analysis of the deleted chromosome segments. The deleted regions in this library encompass 1301 *Bacillus* genes, grouped into 22 COG categories (Figure 3A). Comparing the COGs in deleted and undeleted regions revealed that the carbohydrate metabolism and transport (G) category had the highest proportion of deleted COGs (Figures 3B and S7), suggesting that cumulative deletions in these segments could significantly impact carbohydrate metabolism and transport. Further analysis of the deleted segments showed that each segment contained COGs with diverse functions (Figure 3C). For instance, segments 5 and 67 each had COGs from six functional categories, though the specific categories differed between them. We then analyzed the gene numbers of every category in every chromosome segment to obtain the COG distribution profile in the deleted segments. The probability of having one or two genes belonging to a random COG category in a random segment is 55.7% or 19.8%, respectively (Figure 3D**)**. As examples, the probabilities of having one or two genes in the corresponding COG categories of segment 17 and segment 5 are 100% and 50%, respectively (Figures 3E, 3F, and S8).

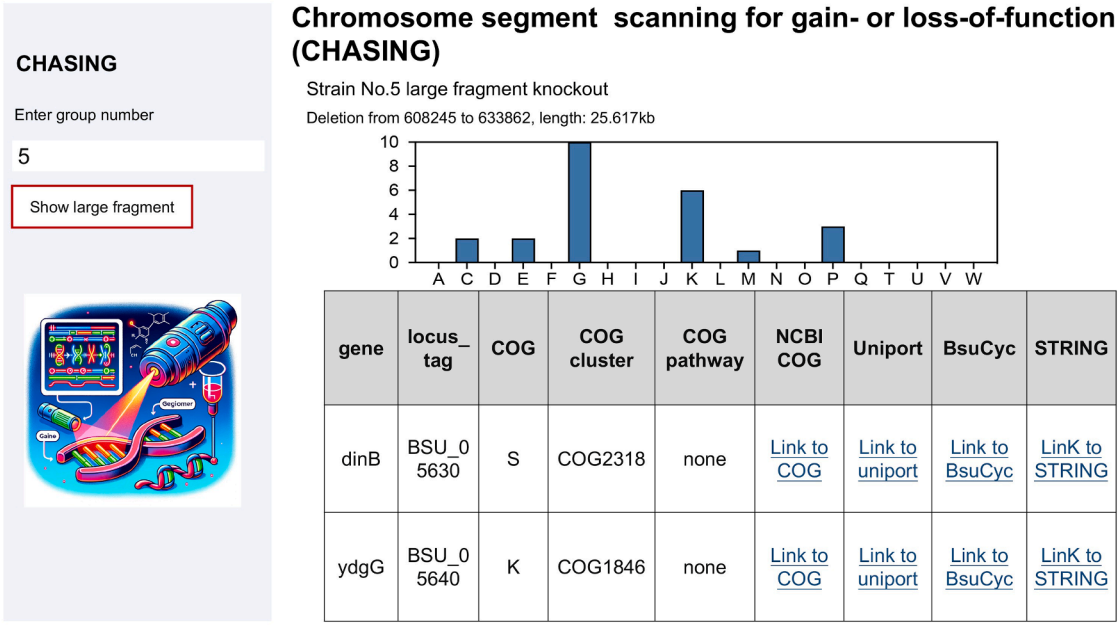
To assist the application of the CHASING strategy in *B. subtilis*, we developed a web tool (https://chasing.cloudmol.org/) to visualize the potential targets of each chromosome segment associated with a particular function. As shown in Figure 3G, the COG distribution pattern along with the deleted genes and the corresponding COG categories would be reported after inputting any segment number of our library. The web tool also presented useful information for each deleted gene by providing the relevant web links, such as the associated metabolism and the protein interaction network. The potential candidates responsible for an interested phenotype could be preliminarily analyzed by dissecting the genes of the corresponding functions. This web tool will highlight the potential of our CSD library and the CHASING strategy in metabolic engineering applications.

Mechanisms underlying the varied cellular fitness

By analyzing the deleted segments of these strains using our web tool, we preliminarily identified the promising targets of sporulation-related COGs in these deletion strains. For example, the deleted segment in strain Δ24 contains *sigE* and *sigF*, in agreement with the previous report that deletion *sigE* and *sigF* severely affected the formation of spores.[22,23] These two genes



G



gene	locus_tag	COG	COG cluster	COG pathway	NCBI COG	Uniport	BsuCyc	STRING
dinB	BSU_05630	S	COG2318	none	Link to COG	Link to uniport	Link to BsuCyc	LinK to STRING
ydgG	BSU_05640	K	COG1846	none	Link to COG	Link to uniport	Link to BsuCyc	LinK to STRING

Figure 3. Bioinformatic analysis of the deleted chromosome segments in *B. subtilis* 168

(A) Characterization of COG categories for all deleted genes in the genome-wide deletion library.

(B) The number of un-deleted and deleted genes in each COG category.

(C) COG distribution pattern in each deleted segment.

(D) Accumulated COG categories with different gene numbers in *B. subtilis* 168. The data was calculated based on the COGs in all deleted chromosome segments.

(E and F) Gene number analysis of each COG category in segments 17 and 5.

(G) The web server of CHASING. The COG distribution profile of each deleted segment of *B. subtilis* was displayed. With the serial number of the chromosome segment as input, the deleted genes along with the corresponding COG category would be presented. The associated metabolism of each deleted gene and the protein interaction network could be obtained by clicking the relevant web links.

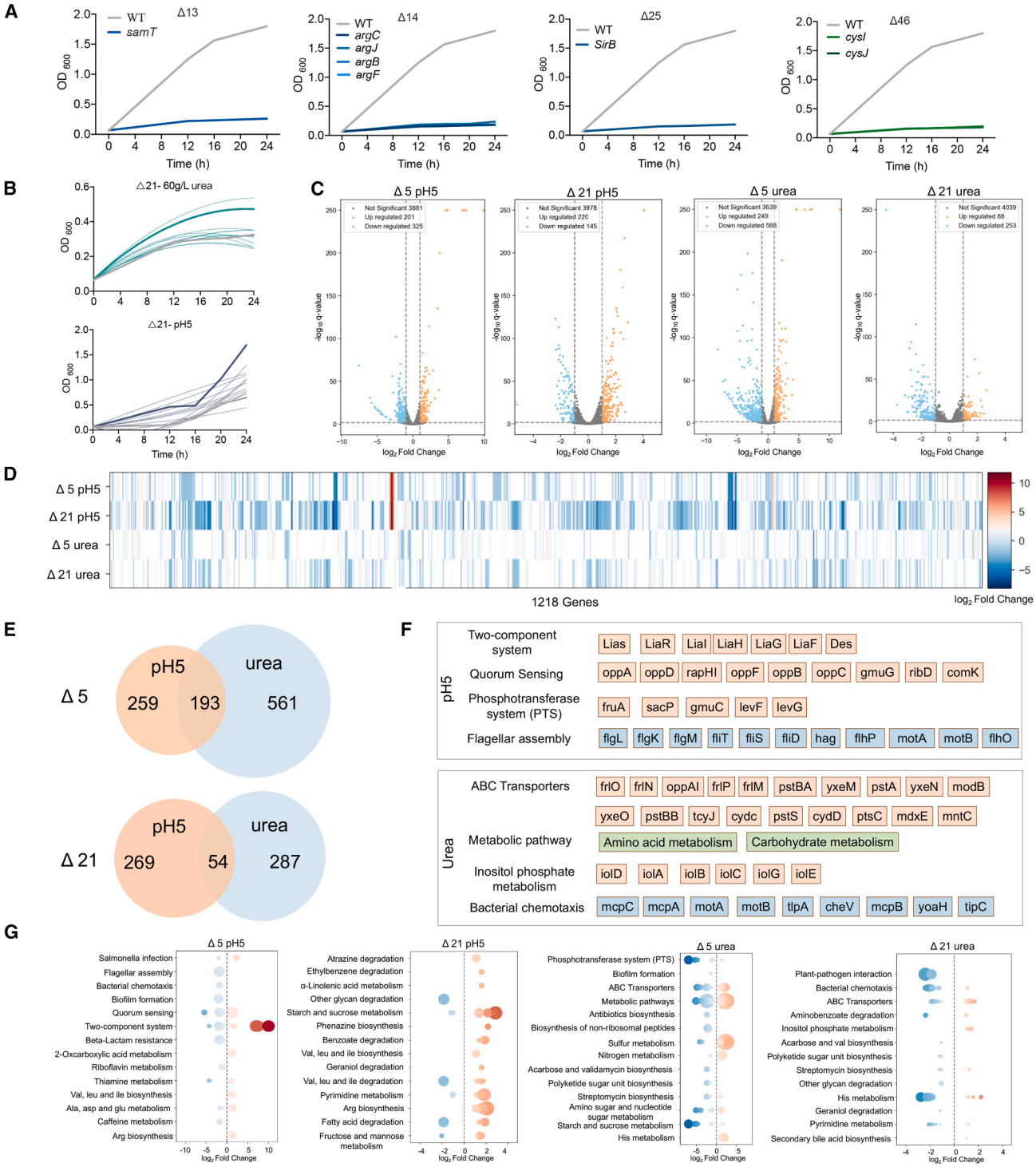


Figure 4. Transcriptome analysis of the strains Δ5 and Δ21 under pH 5 and high urea condition

(A) The growth curves of single-gene-deletion strains in GMII medium. Genes responsible for the cell growth were identified for the strains Δ13, Δ14, Δ25, and Δ46, respectively.

(B) The growth curves of the strain Δ21 and the corresponding single-gene-deletion strains under high urea and pH 5 condition.

(C) Volcano plot of transcriptomic data of the strains Δ5 and Δ21 under pH 5 or high urea conditions. The blue dots represent the downregulated genes, and the orange dots indicate the upregulated ones. The $\log_2$ fold-change in differentially expressed genes was compared to the wild-type *B. subtilis*. The cutoff of the differentially expressed gene was set at a log2-fold change > |1| with a q-value (FDR, padj) < 0.05.

(legend continued on next page)

encode the critical sigma factors that control the transcription of sporulation genes. The deleted segment in Δ5 or Δ15 contains the spore cortex lytic gene *spoL*, and the coat assembly and spore resistance gene *spoVIF*, respectively, likely to be the only genes responsible for the loss-of-sporulation phenotype. The deleted segment in Δ34 contains a sporulation gene cluster, including *spoIIIAA*, *spoIIIAB*, *spoIIIAC*, *spoIIIAD*, *spoIIIAE*, *spoIIIAF*, *spoIIIAG*, and *spoIIIAH*.

To identify conditionally essential genes causing the retarded growth in Δ13, Δ14, Δ25, and Δ46, we analyzed the deleted segments using our web tool. Genes involved in central metabolism were found, including *samT* in Δ13, *argC*, *argJ*, *argB*, and *argF* in Δ14, *sirB* in Δ25, and *cysI* and *cysJ* in Δ46 (Table S1). These genes are involved in the metabolism of homocysteine, arginine, cysteine, and siroheme, respectively. Single-gene knockout experiments confirmed that all eight single-gene knockout strains could not grow in GMII medium, reproducing the phenotypes of the corresponding CSD strains (Figure 4A).

For the strain Δ21, which exhibited a gain-of-growth phenotype, we investigated the deleted segment containing 11 genes (Table S2). Six genes with known functions were unrelated to the phenotype (Figure S9), leaving five unknown genes *ykzR*, *ykvR*, *ykvS*, *ykzS*, and *ykvU* as potential candidates. Deletion of *ykvS* alone led to vigorous growth under 60 g/L urea with a growth rate of 0.123 h^{-1}, 29.5% and 6.03% higher than the wild type (0.095 h^{-1}) and Δ21 (0.116 h^{-1}), respectively (Figure 4B). The other deletions did not restore growth in urea. Deleting the six known-function genes individually confirmed they had no effect on growth in urea. Taken together, this result indicated that the single-gene deletion of *ykvS* in the wild-type strain could reproduce the gain-of-resistance phenotype of strain Δ21 under urea-rich conditions.

It is notable that the Δ21 strain also demonstrated a growth advantage under pH 5 conditions. Since single gene-knockout strains of Δ21 had already been constructed, we measured their growth curves at pH 5. Among them, strain ΔykvQ, lacking a putative sporulation-specific glycosylase gene, showed a similar growth pattern to Δ21, reaching an OD_{600} of 1.26 at 24 h under acidic conditions (Figures 4B and S6B). Additionally, deleting other sporulation-related genes such as *ykvP* and *ykvT*, which encode spore protein and cell wall hydrolase linked to spore cortex-lytic enzymes, also enhanced growth fitness compared to the wild-type. These findings suggest that the growth advantage in strain Δ21 likely stems from the combined deletion of multiple sporulation genes, highlighting the potential of large-scale deletion libraries for discovering robust chassis strains.

To explore the mechanisms underlying chassis robustness, we performed transcriptome analysis, selecting strain Δ5 from PCA group one and strain Δ21 from PCA group two, and identified differentially expressed transcripts under pH 5 and urea-rich conditions (Figures 4C and 4D; Table S1). To isolate the impact of relevant gene deletions, we conducted cluster analysis on Δ5 and Δ21, identifying differentially expressed transcripts and their associated metabolic pathways specific to pH 5 (Figure 4E) and urea-rich conditions (Figure 4F). For the strain Δ5, the most affected metabolism under acid shock is the two-component system and quorum sensing system (Figure 4G), among which the LiaRS system[24] and the Opp system[25] have been reported to be involved in the adaptive responses to the pH stress. These results suggested that the improved fitness in strain Δ5 may be acquired by upregulating the genes of these pathways, leading to a fast and differentiated response under pH 5 conditions. Under urea shock, the most affected metabolism in the strain Δ5 are amino acid and carbohydrate metabolism (Figures 4F and 4G). This could be explained by the feedback inhibition of urea on the production of NH_4^+ and the degradation of amino acids, leading to the rewiring of amino acid and carbohydrate metabolism.[26]

To be noted, the metabolic changes were totally different in the strain Δ21 (Figures 4F and 4G). For example, the most significantly upregulated and downregulated genes under pH 5 conditions are the starch and sucrose metabolism, and glycan degradation pathways, respectively. This could be attributed to the deletion effect of putative sporulation-specific glycosylase gene *ykv*Q, which is responsible for the enhanced tolerance of the strain Δ21 to acid shock (Figure 4B**)**. It has been well documented that glycosylase play important roles in the degradation of glycan.[27] It is possible that the deletion of *ykv*Q resulted in an upregulated level of glycan in Δ21, increasing the abundance of lipopolysaccharide and the toughness of cell wall, thus leading to an improved response to acid stress. Under urea-rich conditions, the most significantly affected genes were associated with the plant-pathogen interaction pathway and His metabolism. Based on this, the absence of hypothetical YkvS protein may enhance the urea tolerance via these pathways.

Apply the genome-scale deletion library to screen for strains with increased plasmid stability

Plasmids could introduce genetic elements into cells, making them a valuable tool for biotechnological applications. We developed a high-throughput system to evaluate plasmid stability using a plasmid with the repA origin and sfGFP reporter. Each strain in the CSD library was transformed with the sfGFP plasmid and cultivated without antibiotic selection for three generations (Figure 5A). After each transfer, fluorescence was measured at 8 h to calculate plasmid preservation rates. After the first transfer, preservation rates ranged from 5

5.8% to 96.8% (Figure 5B). Most strains showed similar fluorescence to the wild-type *B. subtilis*, except Δ5, Δ11, Δ49, and Δ71, which had lower fluorescence. However, after the second

(D) Heatmap of upregulated and downregulated genes under pH5 and high urea conditions.
(E) Venn diagram representation of the differentially expressed transcripts that are specific for pH 5 and high urea condition. The orange and blue colors represent the condition of pH 5 and urea, respectively.
(F) The major pathways with the largest number of differentially expressed genes. The pathways in upregulation and downregulation were indicated in orange and blue color, respectively. The pathways with both upregulated and downregulated genes were indicated in green color.
(G) Scatterplot of differentially expressed genes in KEGG enrichment. Larger size of the dot represents higher number of differentially expressed genes. Deeper color of the dot represents a higher fold-change in differential expressed genes.

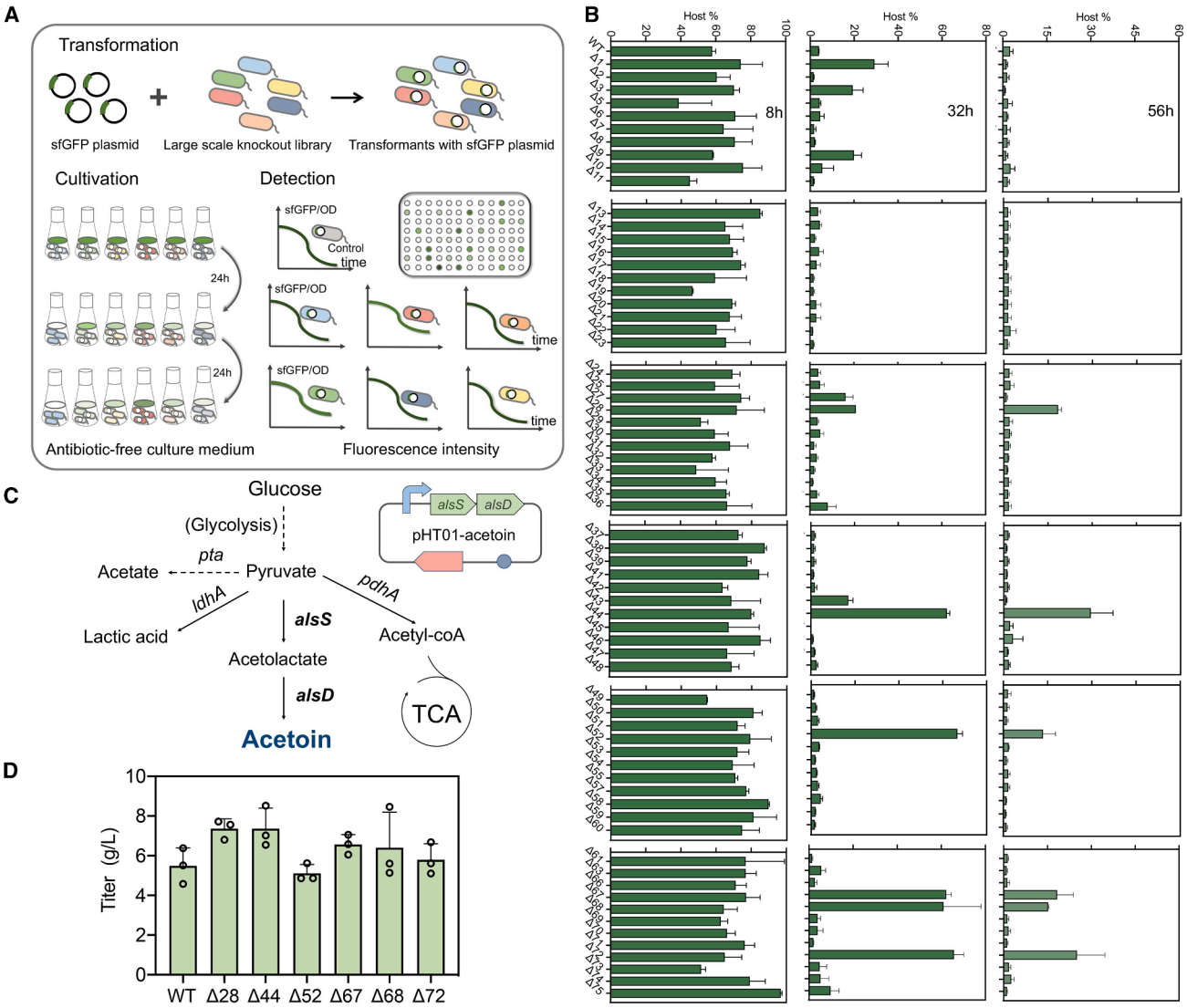


Figure 5. Plasmid stability of the CSD strains under non-selective condition

(A) Schematic diagram of the strain screening for plasmid stability using a high-throughput system. Subculturing of each *sfGFP*-harboring strain was performed every 24 h, and the fluorescence was measured at OD_{474}. Higher fluorescence intensity represents higher plasmid stability.

(B) Plasmid stability analysis of the deletion strains. Sample was collected at 8, 32, and 56 h, respectively, and the fluorescence was measured at OD_{474}.

(C) Metabolic pathway of acetoin production in *B. subtilis*. *alsS*, acetolactate synthase; *alsD*, acetolactate decarboxylase.

(D) Acetoin production of the deletion strains with higher plasmid stability. The fermentation was performed in LB medium without antibiotics. Data are means ± SEM.

transfer, most strains exhibited a significant drop in preservation rates, with the wild-type strain dropping to 4.0%, indicating difficulty in maintaining the plasmid without antibiotics. In contrast, strains Δ1, Δ27, Δ28, Δ44, Δ52, Δ65, Δ66, and Δ70 showed high fluorescence, with an average preservation rate of 48.9%. Even after the third transfer, strains Δ28, Δ44, Δ52, Δ65, Δ66, and Δ70 maintained high plasmid preservation rates of 18.4%, 29.5%, 13.5%, 18.3%, 15.2%, and 25.1%, respectively, identifying them as chassis strains with enhanced plasmid stability. The enhanced plasmid stability in Δ28, Δ52, Δ65, Δ66, and Δ70 strains is the result of the interaction of multiple genomic factors, whereas the improved plasmid stability in the Δ44 strain is due to the deletion of the bacteriophage SPP1 surface receptor gene *yueB*.[28]

To demonstrate the utility of these strains for plasmid maintenance and chemical production, we tested acetoin production in an antibiotic-free system. The *alsS* and *alsD* genes for acetoin biosynthesis were expressed in the plasmid, replacing the sfGFP reporter (Figure 5C). As expected, acetoin production increased in the strains with enhanced plasmid stability. Strains Δ28, Δ44, Δ52, Δ65, Δ66, and Δ70 achieved acetoin titers of 7.36, 7.36, 5.11, 6.56, 6.40, and 5.80 g/L, respectively, surpassing the

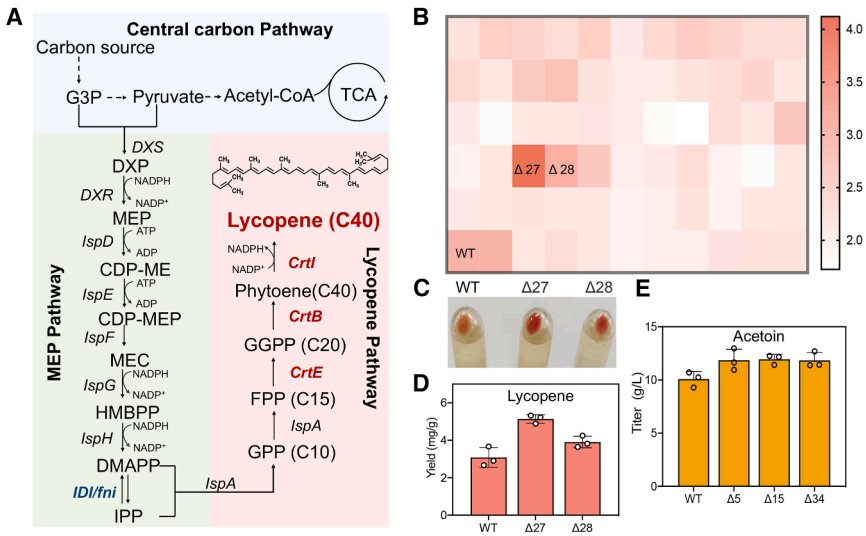


Figure 6. Application of the genome-wide deletion library in metabolic engineering

(A) The engineered pathway of lycopene biosynthesis in *B. subtilis*. The genes in red color are lycopene biosynthetic genes, including the heterologous genes encoding geranylgeranyl diphosphate synthase (CrtE), phytoene synthase (CrtB), and phytoene desaturase (CrtI) from Pantoea ananas.

(B) Heatmap representation of lycopene production in the deletion strains that overexpressing lycopene biosynthetic genes.

(C) The cell pellets of lycopene fermentation samples from the strains Δ27 and Δ28 overexpressing lycopene biosynthetic genes.

(D) Lycopene production in the superior chassis overexpressing lycopene biosynthetic genes and the rate-limiting gene *fni*.

(E) Acetoin production in the sporulation-deficient strains. Data are means ± SEM.

wild-type strain's 5.49 g/L (Figure 5D). These results indicate that chassis strains with high plasmid stability are promising hosts for chemical production.

Apply the genome-scale deletion library to obtain strains with improved chemical production

We selected the food colorant lycopene, known for its antioxidant, anti-aging, and anti-inflammatory properties, as a target product to demonstrate the potential of the genome-scale deletion library in chemical production. While *B. subtilis* could synthesize the intermediates IPP and DMAPP and produce GPP and FPP via *ispA* catalysis, it lacks the complete pathway for lycopene production (Figure 6A). To construct a full lycopene biosynthetic pathway in *B. subtilis*, we cloned and expressed the *crtE*, *crtB*, and *crtI* genes from *Pantoea ananatis*. The lycopene production plasmid expressing *crtEBI* was then transformed into individual strains of the CSD library. After fermentation, culture colors were compared to the wild-type *B. subtilis*, and strains with deeper colors were selected for lycopene extraction and yield analysis. While most mutant strains showed similar color intensity to the control, mutant strains Δ27 and Δ28 exhibited significant lycopene production increases, with yields of 4.12 and 3.24 mg/g dry weight, representing a 47.7% (Figure 6B) and 16.1% (Figure 6C) increase, respectively.

To maximize lycopene production in these chassis strains, we sought to enhance IPP supply by overexpressing biosynthetic enzymes. First, we overexpressed the endogenous rate-limiting enzymes *dxs* and *fni* (a homolog of *idi* from *E. coli*) in the wild-type *B. subtilis*. While overexpressing *dxs* did not significantly affect lycopene production, overexpression of *fni* increased yields by 43.8%. Building on this, we overexpressed *fni* in strains Δ27 and Δ28. Shake-flask fermentation results showed further lycopene yield increases, reaching 5.27 and 4.04 mg/g dry weight, representing 54.1% and 18.2% increases compared to the wild-type strain (Figure 6D).

During *B. subtilis* fermentation, nutrient deficiency triggers sporulation, leading to dormancy and reduced production yield.[29] Inspired by a previous study showing that a Δ*sigE* Δ*sigF* mutant increased acetoin production,[30] we tested chemical production in spore-deficient strains Δ5, Δ15, Δ24, and Δ34 (Figure S3). These strains were transformed with acetoin-producing genes, *alsS* and *alsD*, yielding the engineered acetoin production strains. Fermentation experiments showed that strains Δ5, Δ15, Δ24, and Δ34 produced acetoin at 11.9, 12.0, 12.2, and 11.8 g/L, respectively, compared to 10.1 g/L by the wild type at 72 h (Figure 6E), representing increases of 17.6%, 18.4%, 21.5%, and 17.4%. These results support the trade-off between growth and production.[31]

DISCUSSION

Profiling of arrayed genome-wide deletion or interference libraries lays the foundation for the engineering of robust chassis. However, the shortage of high-throughput libraries for screening the interested phenotype and the limited understanding of the genotype-phenotype relationship hampered this purpose, especially for non-model strains. Machine learning and CRISPR-based technology break the limit in genome editing and increase the feasibility of genome-wide library construction.[32] CRISPR-based long-fragment deletion strategy has been widely reported in diverse organisms, achieving the scarless deletion of genomic fragments up to 134.3 kb, 186.7 kb, and 38 kb in *B. subtilis*,[18] *E. coli*,[33] and *Saccharomyces cerevisiae*,[34,35] respectively. In the present study, we developed AutoGSL tool for the automated design of chromosome-scale knockout libraries. Based on AutoGSL, we employed the CRISPR/Cas9-based long-fragment deletion technique to generate assembled genes knockout strains for phenotyping, successfully establishing an arrayed CSD library covering 31.6% of the *B. subtilis* 168 genome and 32.4% of the non-essential genes. The CRISPR-based assembled genes knockout strategy could also be applied to other microorganisms.

We compared CHASING, Keio collection, and CRISPRi methods in terms of library construction, screening, and

Table 1. Duration and Minimal cost of realizing the phenotype-genotype mapping with a Keio collection, a CRISPRi array, or a CHASING library using the state-of-the-art technologies

Methods		Keio collection	CRISPRi	CHASING
Construction	price	103,700 $	6,150 $	3,500 $
	period	long	short	short
Screening	method	un-restricted	restricted	un-restricted
	process	high waste/complex	low waste/simple	low waste/simple
Application	gene perturbance	completely	not completely	completely
	phenotype stability	stable	unstable	stable

application. For library construction, cost savings were calculated based on primer costs for strain deletion, including N20 amplification, homologous arm, and sequencing. CHASING demonstrated significant cost and time savings compared to the Keio collection in building an arrayed library (Tables 1 and S6). With only 70 strains (excluding controls), the screening scale was well-suited for 96-well plates and easy to handle. Additionally, CHASING supports pooled screening for specific applications by using chromosome segment-specific PCR primers for the rapid identification of deletions. This enhances versatility and efficiency. CHASING's cost savings are comparable to CRISPRi when considering genome coverage (Table 1), with key advantages such as unrestricted screening, phenotypically stable strains, and precise gene expression perturbation. Unlike pooled CRISPRi libraries, the arrayed CSD library allows for single-pot assays, providing accurate readouts in experiments such as growth measurements and morphology phenotyping.

To assist the genotyping of interested phenotypes, we designed a user-friendly web server to scan the COG distribution profile of the deleted chromosome segment in *B. subtilis*, and showed that the probability to link a single gene directly to an observed phenotype is 55.7% for our obtained library. Regarding the simplified COG category, some functions such as stress resistance, chemical production, or sporulation are not assigned to any specific category of COG family, leading the corresponding genes be shadowed using the CHASING strategy. This could be addressed by generating a PCOF with more diversified categories by incorporating more protein databases such as the InterPro database (https://www.ebi.ac.uk/interpro/), into the PCOF. Another challenge is the huge number of hypothetical and unknown genes in the genome, and especially the difficulty in annotating them using homology-based approaches. With the explosion of protein sequences and the maturity of AlphaFold database, this might be addressed by leveraging the machine learning revolution in protein bioinformatics.[36] For every chromosome segment, a more specified categorization generated a higher probability of linking one gene to an interested phenotype. Therefore, the advance of cutting-edge technologies and the advent of interactive resources give promise into employing CHASING strategy for rapid genotyping. As an available accessory, our web server could be further upgraded by incorporating the renewed PCOF.

The *B. subtilis* genome could be divided into 148 non-essential gene segments (Figure S10A), each assigned to a specific COG. For instance, segment 82 contains 11 COG categories, and segment 31 contains 15 (Figure S10B). We analyzed gene numbers per category in each segment, finding a 46.1% and 20.6% probability of one or two genes belonging to a random COG in a segment, respectively (Figure S10C). Thus, approximately 34% of cumulative COG categories contain more than two genes, some forming gene clusters. For example, eight COG categories in segments 82 or 104 contain only one or two genes, whereas 30 genes in segment 31 belong to COG K, and 27 genes in segment 62 belong to COG N, suggesting gene clusters in larger segments. Deletion of segments 31 or 62 increases the likelihood of observing K- or N-related phenotypes, indicating that the robust phenotype of a strain may not be reproduced through single-gene deletion due to the synergistic effects of multiple genes. For instance, strain Δ21 under pH 5 conditions could not be replicated by single-gene knockouts. This robustness may be missed in single-gene perturbation libraries due to gene redundancy. The CHASING strategy, therefore, has unique potential in screening robust chassis.

Eliminating a large quantity of non-essential genes affect the transcription and translation in the cell, thus adjusting the resource allocation and reducing the metabolic complexity for diverse applications.[37–39] The complicated network interconnections lay the foundation for the robustness of the CSD library, thereby the targeted phenotypes would not be shadowed by the knockout of unrelated genes, allowing mutants of interest to be screened out (Figure 1E). In this regard, our library provided versatile resources for yielding diverse mutants, offering promise to gain more comprehensive knowledge toward cellular life and interactive metabolism beyond the currently existing genome-minimized strains.

The applicability of the CHASING strategy was further evaluated by analyzing the genomes of *E. coli* and *Bacillus thuringiensis*. For *E. coli*, the whole genome could be divided into 184 large segments by essential genes (Figure S11). The probability of having one or two genes belonging to a random COG category in a random segment is 45.8% or 20.8%, respectively (Figure S10D). For *B. thuringiensis*, the whole genome could be divided into 437 large segments by essential genes (Figure S12). The probability of having one or two genes belongs to a random COG category in a random segment is 64.6% or 19.8%, respectively (Figure S10E). These results further demonstrated the applicability and generality of the CHASING strategy.

We developed AutoGSL to construct a systematic arrayed library via assembled gene knockouts. An arrayed genome-wide library was established in *B. subtilis*, enabling the identification of strains with stress tolerance, enhanced plasmid stability, sporulation defects, and improved chemical production. Our

platform proved effective for genome-wide perturbation and high-throughput functional screening. The CHASING strategy showed potential in screening for desired phenotypes and elucidating genotype-phenotype relationships. Additionally, we introduced the COG Transformer encoding method to train machine learning models for predicting growth phenotypes, offering a promising approach for large-scale genomic feature extraction.

Limitations of the study

In this study, a genome-scale library was constructed to effectively screen for phenotypes associated with both loss-of-function and gain-of-function mutations. However, due to the presence of numerous knocked-out genes within the phenotypic profiles, pinpointing the specific gene or a few genes responsible for the observed phenotype remains challenging. Therefore, a more refined categorization of COG functional classifications is required to facilitate the rapid identification of phenotype-associated genes. Furthermore, our method is theoretically applicable across various species, and could potentially be expanded for use in a broader range of organisms in the future.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Shuyuan Guo (guosy@bit.edu.cn).

Materials availability

Plasmids generated in this study are available upon request.

Data and code availability

- The raw sequencing data processed by this study was deposited in the publicly accessible database Sequence Read Archive (SRA) with accession numbers PRJNA1031996.
- All original code has been deposited on GitHub at https://github.com/YanXia157/COG-Transformer.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2024YFA0917500), the Natural Science Foundation of China (grant number 32370095), and the Hebei Natural Science Foundation (grant number C2023105022). Part of the experiments were carried out in the Biological & Medical Engineering Core Facilities of the Beijing Institute of Technology.

AUTHOR CONTRIBUTIONS

Y.-X.H. generate the concept. S.Y.G., Y.-X.H., and Y.X. conceptualized the project. Y.X., Z.Y.L., Z.R.H., J.L., and P.Y.D. carried out the experiments. Y. X. completed bioinformatics analysis and machine learning models. S.Y.G., Y.-X.H., Y.X., and L.C.S. analyzed the experiments and transcriptome data and wrote the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112484.

Received: December 3, 2024
Revised: January 23, 2025
Accepted: April 15, 2025
Published: April 17, 2025

REFERENCES

1. de Vienne, D., and Capy, P. (2022). Special issue on "The relationship between genotype and phenotype: new insight into an old question". Genetica *150*, 151.
2. Brochado, A.R., and Typas, A. (2013). High-throughput approaches to understanding gene function and mapping network architecture in bacteria. Curr. Opin. Microbiol. *16*, 199–206.
3. van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods *6*, 767–772.
4. Kazi, M.I., Schargel, R.D., and Boll, J.M. (2020). Generating Transposon Insertion Libraries in Gram-Negative Bacteria for High-Throughput Sequencing. J. Vis. Exp. https://doi.org/10.3791/61612.
5. Wang, T., Guan, C., Guo, J., Liu, B., Wu, Y., Xie, Z., Zhang, C., and Xing, X. H. (2018). Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. Nat. Commun. *9*, 2475.
6. Rousset, F., Cui, L., Siouve, E., Becavin, C., Depardieu, F., and Bikard, D. (2018). Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. PLoS Genet. *14*, e1007749.
7. Yao, L., Shabestary, K., Björk, S.M., Asplund-Samuelsson, J., Joensson, H.N., Jahn, M., and Hudson, E.P. (2020). Pooled CRISPRi screening of the cyanobacterium Synechocystis sp PCC 6803 for enhanced industrial phenotypes. Nat. Commun. *11*, 1666.
8. Shin, J., Bae, J., Lee, H., Kang, S., Jin, S., Song, Y., Cho, S., and Cho, B.K. (2023). Genome-wide CRISPRi screen identifies enhanced autolithotrophic phenotypes in acetogenic bacterium Eubacterium limosum. Proc. Natl. Acad. Sci. USA *120*, e2216244120.
9. de Bakker, V., Liu, X., Bravo, A.M., and Veening, J.W. (2022). CRISPRi-seq for genome-wide fitness quantification in bacteria. Nat. Protoc. *17*, 252–281.
10. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191.

11. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

12. Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., et al. (2009). Update on the Keio collection of Escherichia coli single-gene deletion mutants. Mol. Syst. Biol. *5*, 335.

13. Koo, B.M., Kritikos, G., Farelli, J.D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J.M., et al. (2017). Construction and Analysis of Two Genome-Scale Deletion Libraries for Bacillus subtilis. Cell Syst. *4*, 291–305.e7.

14. Puddu, F., Herzog, M., Selivanova, A., Wang, S., Zhu, J., Klein-Lavi, S., Gordon, M., Meirman, R., Millan-Zambrano, G., Ayestaran, I., et al. (2019). Genome architecture and stability in the Saccharomyces cerevisiae knockout collection. Nature *573*, 416–420.

15. Peters, J.M., Colavin, A., Shi, H., Czarny, T.L., Larson, M.H., Wong, S., Hawkins, J.S., Lu, C.H.S., Koo, B.M., Marta, E., et al. (2016). A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. Cell *165*, 1493–1506.

16. Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., and Andrews, B. (2019). Global Genetic Networks and the Genotype-to-Phenotype Relationship. Cell *177*, 85–100.

17. Henser-Brownhill, T., Monserrat, J., and Scaffidi, P. (2017). Generation of an arrayed CRISPR-Cas9 library targeting epigenetic regulators: from high-content screens to in vivo assays. Epigenetics *12*, 1065–1075.

18. Tian, J., Xing, B., Li, M., Xu, C., Huo, Y.X., and Guo, S. (2022). Efficient Large-Scale and Scarless Genome Engineering Enables the Construction and Screening of Bacillus subtilis Biofuel Overproducers. Int. J. Mol. Sci. *23*, 4853.

19. Sanderson, T., Bileschi, M.L., Belanger, D., and Colwell, L.J. (2023). ProteInfer, deep neural networks for protein functional inference. Elife *12*, e80942.

20. Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. Science *379*, 1358–1363.

21. Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D., and Koonin, E.V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. *49*, D274–D281.

22. Ju, J., Luo, T., and Haldenwang, W.G. (1998). Forespore expression and processing of the SigE transcription factor in wild-type and mutant Bacillus subtilis. J. Bacteriol. *180*, 1673–1681.

23. Rodríguez Ayala, F., Bartolini, M., and Grau, R. (2020). The stress-responsive alternative sigma factor SigB of Bacillus subtilis and its relatives: an old friend with new functions. Front. Microbiol. *11*, 1761.

24. Jordan, S., Hutchings, M.I., and Mascher, T. (2008). Cell envelope stress response in Gram-positive bacteria. FEMS Microbiol. Rev. *32*, 107–146.

25. Liu, W., Huang, L., Su, Y., Qin, Y., Zhao, L., and Yan, Q. (2017). Contributions of the oligopeptide permeases in multistep of Vibrio alginolyticus pathogenesis. Microbiologyopen *6*, e00511.

26. Chandel, N.S. (2021). Amino Acid Metabolism. Cold Spring Harbor Perspect. Biol. *13*, a040584.

27. A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, and M.E. Etzler, eds. (2009). Essentials of Glycobiology (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press Copyright © 2009, The Consortium of Glycobiology).

28. Guo, Y., Xia, Y., Liang, Z., Yang, S., Guo, S., Sun, L., and Huo, Y.-X. (2024). Plasmid-stabilizing strains for antibiotic-free chemical fermentation. ACS Synth. Biol. *13*, 2820–2832.

29. Klausmann, P., Hennemann, K., Hoffmann, M., Treinen, C., Aschern, M., Lilge, L., Morabbi Heravi, K., Henkel, M., and Hausmann, R. (2021). Bacillus subtilis High Cell Density Fermentation Using a Sporulation-Deficient Strain for the Production of Surfactin. Appl. Microbiol. Biotechnol. *105*, 4141–4151.

30. Wang, Q., Zhang, X., Ren, K., Han, R., Lu, R., Bao, T., Pan, X., Yang, T., Xu, M., and Rao, Z. (2022). Acetoin production from lignocellulosic biomass hydrolysates with a modular metabolic engineering system in Bacillus subtilis. Biotechnol. Biofuels Bioprod. *15*, 87.

31. Zhu, M., Wang, Q., Mu, H., Han, F., Wang, Y., and Dai, X. (2023). A fitness trade-off between growth and survival governed by Spo0A-mediated proteome allocation constraints in Bacillus subtilis. Sci. Adv. *9*, eadg9733.

32. Bock, C., Datlinger, P., Chardon, F., Coelho, M.A., Dong, M.B., Lawson, K.A., Lu, T., Maroc, L., Norman, T.M., Song, B., et al. (2022). High-content CRISPR screening. Nat. Rev. Methods Primers *2*, 9.

33. Huang, C., Guo, L., Wang, J., Wang, N., and Huo, Y.X. (2020). Efficient long fragment editing technique enables large-scale and scarless bacterial genome engineering. Appl. Microbiol. Biotechnol. *104*, 7943–7956.

34. Hao, H., Wang, X., Jia, H., Yu, M., Zhang, X., Tang, H., and Zhang, L. (2016). Large fragment deletion using a CRISPR/Cas9 system in Saccharomyces cerevisiae. Anal. Biochem. *509*, 118–123.

35. Li, Z.H., Liu, M., Lyu, X.M., Wang, F.Q., and Wei, D.Z. (2018). CRISPR/Cpf1 facilitated large fragment deletion in Saccharomyces cerevisiae. J. Basic Microbiol. *58*, 1100–1104.

36. Durairaj, J., Waterhouse, A.M., Mets, T., Brodiazhenko, T., Abdullah, M., Studer, G., Tauriello, G., Akdel, M., Andreeva, A., Bateman, A., et al. (2023). Uncovering new families and folds in the natural protein universe. Nature *622*, 646–653.

37. Reuß, D.R., Altenbuchner, J., Mäder, U., Rath, H., Ischebeck, T., Sappa, P.K., Thürmer, A., Guérin, C., Nicolas, P., Steil, L., et al. (2017). Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism. Genome Res. *27*, 289–299.

38. Kim, S.J., and Oh, M.K. (2023). Minicell-forming Escherichia coli mutant with increased chemical production capacity and tolerance to toxic compounds. Bioresour. Technol. *371*, 128586.

39. Zhang, X., He, A., Zong, Y., Tian, H., Zhang, Z., Zhao, K., Xu, X., and Chen, H. (2023). Improvement of protein production in baculovirus expression vector system by removing a total of 10 kb of nonessential fragments from Autographa californica multiple nucleopolyhedrovirus genome. Front. Microbiol. *14*, 1171500.

40. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646.

41. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.20.500902.

42. Peng, C., Zhu, S., Lu, J., Hu, X., and Ren, L. (2020). Transcriptomic analysis of gene expression of menaquinone-7 in Bacillus subtilis natto toward different oxygen supply. Food Res. Int. *137*, 109700.

43. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
LB Broth Powder	Solarbio	Cat# L1010
Agar	Solarbio	Cat# A8190
Agarose	Tsingke	Cat# TSJ001
IPTG	Macklin	Cat# I811719
Tryptone	Solarbio	Cat# T8490
Yeast extract	Solarbio	Cat# Y8020
NaCl	Solarbio	Cat# S8210
KCl	Solarbio	Cat# P9921
$MgCl_2$	Solarbio	Cat# M8161
$MgSO_4$	Solarbio	Cat# M9400
Glucose	Solarbio	Cat# G8150
Kanamycin	Macklin	Cat# 768526
Chloromycetin	Macklin	Cat# C6200
Ampicillin	Macklin	Cat# A6265
5 × M9 Minimal Salt	Topbiol	Cat# M1076
Vitamin B1	Macklin	Cat# A832636
PBS, 0.01M pH 7.2–7.4	Solarbio	Cat# P1020
DpnI	ABclonal	Cat# RK21109
SOC Medium	Macklin	Cat# S917758
Critical commercial assays		
2 × Phanta Flash Master Mix	Vazyme	Cat# P510-01
2 × Taq Master Mix (Dye Plus)	Vazyme	Cat# P112-01
FastPure Gel DNA Extraction Mini Kit	Vazyme	Cat# DC301-01
FastPure Plasmid Mini Kit	Vazyme	Cat# DC201-01
2 × MultiF Seamless Assembly Mix	ABclonal	Cat# RK21020
Deposited data		
Raw and processed RNA sequencing data	This paper	SRA: PRJNA1031996
Experimental models: Organisms/strains		
Escherichia coli JM109	Lab stock	N/A
Bacillus subtilis 168	Lab stock	N/A
Bacillus subtilis 168 konck-out strainsΔ1-Δ75	This paper	N/A
Oligonucleotides		
Gene editing verification primers	This paper	Table S5
Vector construction primers	This paper	Table S5
Recombinant DNA		
Plasmids generated in this study	This paper	Table S4
Software and algorithms		
GraphPad Prism 8	GraphPad Software	http://www.graphpad.com/
Snap Gene Viewer (v6.2.2)	Snap Gene software	http://www.snapgene.com/
CRISPR WebServer	Opensource	https://zlab.squarespace.com/guide-design-resources
HTseq software (V 0.6.1)	Opensource	https://bioweb.pasteur.fr/packages/pack@HTSeq@0.6.1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R (v3.43)	Opensource	https://www.r-project.org/
Flow Jo (V10.8.1)	BD Biosciences	https://www.flowjo.com/
COG Transformer	This paper	https://github.com/YanXia157/COG-Transformer

METHOD DETAILS

Plasmid construction for chromosome segment deletion

The relevant plasmids and primers can be found in Tables S3 and S4. All primers were synthesized by Genewiz (Tianjin, China). The knockout plasmid backbone, p8999-2N20, used in our study, contains Cas9 under mannose-induced promoters and two sgRNA expression cassettes driven by the constitutive promoter P_{veg}. To target the chromosome segment, we carefully selected two highly efficient N20 sequences located at the beginning and end of the target region by AutoGSL. For the rapid and simple construction of various knockout plasmids, we simultaneously introduced two sgRNAs using Gibson assembly. Next, the plasmid containing the target sgRNA serves as the template for the second round of PCR. The upstream and downstream homology arm sequences are amplified from the *B. subtilis* genome, and then fragments are generated using overlap extension PCR (OEPCR). Subsequently, the backbone and fragments are joined together through Gibson assembly, resulting in the formation of a complete knockout plasmid.

Construction of genome-scale chromosome-segment deletion library

In order to construct a genome-scale deletion library for *B. subtilis*, we selected the N20 sequences to target the proposed regions and constructed the plasmids using the highly efficient sgRNAs along with 2 kb repair template. The resulting plasmids were transformed into *B. subtilis* 168 by natural transformation method. 20 μL of cells was plated on the LB agar plates with 20 μg/mL kanamycin and 0.2% mannose, and incubated at 30°C for editing. After 15 h, we selected 8 colonies from each plate for deletion validation. Two pairs of validation primers were designed for each deleted segment, the first pair was designed inside the deleted segment and the second pair was designed outside the region of homology arm. Theoretically, the positive colony would yield PCR products with the correct length using the first pair and no product using the second pair. The editing efficiency was calculated as the ratio of the number of positive colonies to the selected ones. The results were further validated through Sanger sequencing. Once the single colony that containing the deleted segment was verified, plasmid elimination was carried out using the method as previously.[18] The chromosome-segment-deletion strains listed in Table S5 were preserved at -80°C.

Bioinformatics analysis of the genome-scale deletion library

COG annotation was accomplished using eggnog-emapper. All available data of protein-protein interactions in *B. subtilis* were retrieved from the STRING database.[40] Data of protein-protein interactions with a combined score greater than 800 were considered reliable. Python was employed to generate the graph of protein-protein interaction while NetworkX and Matplotlib was used for graph construction and visualization, respectively. The distribution of undeleted genes and deleted genes was embedded using a 2D representation by the t-SNE projections.

COG transformer model architecture

We developed the COG Transformer, a model based on the ESM2 architecture,[41] with six transformer layers, an embedding dimension of 320, and 20 attention heads. The input to the model consists of protein sequences encoded using a vocabulary derived from COG functional categories, alongside regular tokens (<pad>, <mask>, <cls>, <eos>, <sep>, and <unk>). The total vocabulary size for the model was 33 tokens.

Dataset and preprocessing

The input dataset was based on functional annotations from the EggNOG database, which was split into training, validation, and test sets using an 80-10-10 ratio. The COG sequences were tokenized into chunks of up to 512 tokens using the provided COG vocabulary, with each sequence wrapped in a <cls> and <eos> token. We applied dynamic token masking during training, where 15% of tokens were masked, following the distribution: 80% masked (<mask> token), 10% replaced with a random token, and 10% unchanged.

Masked language model (MLM) training objective

The COG Transformer was trained using the MLM objective, where the goal is to predict masked tokens in the input sequence. For a given sequence of tokens $x = \{x_1, x_2, \ldots x_n\}$, a subset of the tokens $\{x_{i1}, x_{i2}, \ldots x_{in}\}$ is replaced with the <mask> token.

The objective is to maximize the likelihood of predicting the correct tokens at these masked positions. Formally, the MLM objective can be written as:

$$\mathcal{L}_{MLM} = -\sum_{i \in M} logP(x_i | x_{masked})$$

where M is the set of masked indices and $P(x_i | x_{masked})$ is the probability of predicting the correct token x_i at the i -th position given the masked input sequence.

Training procedure

The COG Transformer was trained using the MLM objective described above. The total number of epochs was set to 50, with mini-batches of size 16. After each epoch, validation loss was calculated to track generalization. The model achieving the lowest validation loss was selected as the best-performing model and saved for subsequent testing.

Phenotype screening

The survival capability of the resulting 75 strains were examined in various conditions, including LB medium, GMII medium, GMII medium under acidic conditions (pH 5), and GMII medium with 60 g/L urea. The strains in the library were streaked onto LB agar plates and incubated at 37°C for 12 h. Then, three single colonies were picked from each plate and inoculated into the fresh culture, followed by incubation at 37°C and 220 rpm for 12 h. Subsequently, they were transferred into 20 mL of LB broth in shake flasks at a 1% of inoculation ratio and incubated at 37°C and 220 rpm for 24 h. During the incubation period, growth measurements were taken every 4 h. For growth measurement, 200 μL of the culture was transferred into a 96-well plate, and the OD_{600} was measured using a spectrophotometer. Blank LB medium was used as a control, and the OD_{600} values were obtained by subtracting the blank control values. Growth curve of each strain was plotted based on these values.

Transcriptomic analysis

The cell cultures were collected in logarithmic phase (OD_{600} of 1.0) by centrifugation at 5,000 rpm for 10 min under 4°C. The cell cultures were then subjected to liquid nitrogen frozen and stored at –80°C. Total RNA was extracted from the collected cells using the methods reported previously.[42] The library construction, purification, and Illumina sequencing were completed by GENEWIZ company (Tianjin, China). To ensure the quality of sequencing data, thorough pre-processing of the raw data was conducted, including the meticulous filtering of low-quality data, meticulous removal of contaminants, and precise trimming of adapter sequences. The gene expression was analyzed with the HTseq software (V 0.6.1) using the FPKM (Fragments Per Kilobase per Million reads) method, which was proposed by Mortazavi in 2008.[43] The differential expression analysis was conducted using the DESeq2 (V1.6.3) package from Bioconductor. The cutoff of the differentially expressed gene was set at a log2-fold change > |1| with a q-value (FDR, padj) < 0.05.

Screening for plasmid stability

We constructed the plasmid pXY-sfGFP and transformed it into each strain of the genome-scale deletion library. The resulting strains were inoculated into 5 mL of LB liquid medium that containing antibiotics and cultured at 37°C and 200 rpm for 8 h. Subsequently, the seed culture was transferred to LB medium without antibiotics at a inoculation ratio of 1%. Subculturing was performed every 24 h and 200 μL samples were taken at 8 h, 32 h, and 56 h for fluorescence measurement at OD_{474}. The plasmid stability was calculated using the formula: Plasmid stability = (Fluorescence value in LB without antibiotics / Fluorescence value in LB with antibiotics) ×100%.

Acetoin fermentation and product detection

The acetoin fermentation plasmid was obtained from our lab stock. Strains of interested were transformed with these plasmids. Single colonies were selected and inoculated into 5 mL of LB liquid medium for incubation at 30°C and 200 rpm for 8 h. Subsequently, the seed culture was transferred to 20 mL fermentation medium with or without antibiotic as required. The fermentation process was conducted at 30°C and 200 rpm, and the sample was taken every 24 h. To detect the acetoin product, gas chromatography was performed using an instrument (PANNA GCA91) equipped with an FID detector. All experiments were performed in triplicates, and the data was presented as an average value.

Screening for lycopene production

Lycopene biosynthetic genes *crtE*, *crtB*, and *crtI* from *Pantoea ananas* were expressed in plasmid pHT01, yielding strain pHT01-crtEBI. Seed culture was inoculated at 5% into 20 mL LB medium with antibiotics, and fermentation was conducted in 50 mL shake flasks. Samples were taken every 24 h for cell dry weight and lycopene production. After centrifugation, the pellets were dried at 65°C to calculate lycopene production per unit dry weight. Lycopene was extracted using acetone: pellets were resuspended in ddH_2O and lysed with beads for 20 min. Lycopene extraction was conducted at 55°C for 1 h, and the titer was determined by measuring absorbance at 474 nm.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses are reported as mean ± standard error of the mean (s.e.m.), calculated as $\sigma/\sqrt{n}$, where σ represents the standard deviation and n denotes the number of biological replicates (n = 3 independent experiments). Error bars indicating the s.e.m. are shown in Figures 5B, 5D, 6D, and 6E. All analyses were performed using GraphPad Prism, Ver. 8 (GraphPad Software, Boston, MA).