Pooled CRISPR interference screens enable high-throughput functional genomics study and elucidate new rules for guide RNA library design in *Escherichia coli*

Tianmin Wang[a], Jiahui Guo[a], Changge Guan[a], Yinan Wu[a], Bing Liu[b], Zhen Xie[c,d], Chong Zhang[a,d]#, Xin-Hui Xing[a,d]

[a]MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

[b]Beijing Syngentech Co., Ltd., Beijing 102206, China

[c]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and System Biology, Department of Automation, Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing 100084, China

[d]Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

Running Head: CRISPRi pooled screen in *E. coli*

#Address correspondence to Chong Zhang, chongzhang@tsinghua.edu.cn

Abstract: Clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9 technology provides potential advantages in high-throughput functional genomics analysis in prokaryotes over previously established platforms based on recombineering or transposon mutagenesis. In this work, as a proof-of-concept to adopt CRISPR/Cas9 method as a pooled functional genomics analysis platform in prokaryotes, we developed a CRISPR interference (CRISPRi) library consisting of 3,148 single guide RNAs (sgRNAs) targeting the open reading frame (ORF) of 67 genes with known knockout phenotypes and performed pooled screens under two stressed conditions (minimal and acidic medium) in *Escherichia coli*. Our approach confirmed most of previously described gene-phenotype associations while maintaining < 5% false positive rate, suggesting that CRISPRi screen is both sensitive and specific. Our data also supported the ability of this method to narrow down the candidate gene pool when studying operons, a unique structure in prokaryotic genome. Meanwhile, assessment of multiple loci across treatments enables us to extract several guidelines for sgRNA design for such pooled functional genomics screen. For instance, sgRNAs locating at the first 5% upstream region within ORF exhibit enhanced activity and 10 sgRNAs per gene is suggested to be enough for robust identification of gene-phenotype associations. We also optimized the hit-gene calling algorithm to identify target genes more robustly with even fewer sgRNAs. This work showed that CRISPRi could be adopted as a powerful functional genomics analysis tool in prokaryotes and provided the first guideline for the construction of sgRNA libraries in such applications.

Importance: To fully exploit the valuable resource of explosive sequenced microbial genomes, high-throughput experimental platform is needed to associate genes and phenotypes at the genome level, giving microbiologists the insight about the genetic structure and physiology of a microorganism. In this work, we adopted CRISPR interference method as a pooled high-throughput functional genomics platform in prokaryotes with *Escherichia coli* as the model organism. Our data suggested that this method was highly sensitive and specific to map genes with previously known phenotypes, potent to act as a new strategy for high-throughput microbial genetics study with advantages over previously established methods. We also provided the first guideline for the sgRNA library design by comprehensive analysis of the screen data. The concept, gRNA library design rules and open-source scripts of this work should benefit prokaryotic genetics community to apply high-throughput mapping of defined gene set with phenotypes in a broad spectrum of microorganisms.

Introduction

Experimental approaches for gene-phenotype mapping are needed to keep up with the pace of explosive microbial genome sequencing capacity. To handle thousands of poorly characterized genes in microbes, such experimental methods are expected to be high-throughput, enabling profiling of genome-wide gene set at multiple growth conditions simultaneously. One promising method is the high-throughput pooled functional genomics analysis, commonly performed by mixing a large number of mutants and monitoring their abundance change during growth with next-generation sequencing(1–4).

Two categories of methods have been established for such purposes. One widely applied approach depends on random transposon derived gene knockout library(1, 2, 5). This method suffers from the random insertion of transposons into the chromosome, resulting in the bias towards genes with long coding region as well as only applicable to genome-wide rather than focused library. Moreover, complicated enzymatic and purification steps are required for library preparation based on transposon mutagenesis(1, 2). An alternative approach depends on quantifying DNA barcodes in a pooled format. The DNA barcodes are previously associated with the mutations introduced by recombineering either in a pooled (such as trackable multiplex recombineering (TRMR)(6)) or arrayed (such as the *Saccharomyces cerevisiae*

deletion collection(4), where DNA barcodes were incorporated into each deletion strain) library. Such strategy also faces problems when trying to apply to broader spectrum of microorganisms, because recombineering system is only established in a limited number of microorganisms. To address these issues, we need a novel platform for high-throughput pooled functional genomics analysis in microorganisms, where the molecular mechanism should be general to as many hosts as possible, the library should be tailored made to either genome-wide or focused gene set and the bias issue should be avoided as much as possible.

Recently developed CRIPSR/Cas9 technology can be used for versatile genome editing guided by a programmed sgRNA in many organisms(7–11). Efforts along this line has resulted in many sgRNA libraries to induce genome-wide gene knockout(12–14), knockdown(15, 16) and activation(15, 17) for functional genomics screens. However, to the best of our knowledge, all of these works have so far focused on eukaryotic organisms, especially in mammalian cell lines. Compared with abovementioned pooled functional genomics platforms for microorganisms that have been previously established, CRISPR/Cas9 system provides several advantages. Firstly, CRISPR/(d)Cas9 activity has been confirmed in diverse prokaryotes(16, 18–23) and thus provides a more broadly adopted platform than recombineering. Secondly, the target-specificity-coding region in sgRNA consists of only ~20 nucleotides, compatible with massively parallel microarray oligonucleotide synthesis and next generation sequencing. This

makes it very easy to prepare either the tailor made library for any defined gene set via microarray oligonucleotide synthesis or the sequencing library using a simple PCR reaction. Thirdly, because the CRISPRi complex is a recombineering-free *trans*-regulatory element and sgRNA library is constructed in the plasmid format, such library can be readily transformed into any number of hosts with different genetic backgrounds in a relatively more unbiased manner.

Considering these advantages and the absence of its application in the field of high-throughput pooled functional genomics in prokaryotes, we sought to establish CRISPR interference (CRISPRi) system, a CRISPR/Cas9 deriative using a nuclease-activity-free Cas9 mutant protein to repress transcription, as a pooled functional genomics study platform in prokaryotic organisms. It should be noted that the fundamental differences between eukaryotic and prokaryotic genomes as well as their transcription regulation(24) increase the risk to directly applying the rules drawn from previous eukaryotic library design as well as screen experience to bacteria or archaea. For example, chromatin accessibility and nucleosome occupancy, which are unique structures in eukaryotic genomes(25), were found to have significant impact on sgRNA activity(26–28). On the other hand, in current CRISPRi sgRNA library design guideline in eukaryotic cells, the target site is selected around the transcription start site (TSS)(26). However, many genes in prokaryotic genomes are organized in operons

co-transcribed as polycistronic mRNA, where a common promoter drives the transcription of all these genes. In these cases, directing CRISPRi complex to the promoter region is expected to repress the transcription of the whole operon, rendering the method fail to identify the contributing individual gene responsible for the studied phenotype.

To address these issues, informed by existing knockout-phenotype data, we designed a synthetic sgRNA library consisting of 3,148 sgRNAs (including 400 control sgRNAs) targeting 67 genes with known phenotype and performed a proof-of-concept screen in *E. coli*. Facing the genome structure difference for prokaryotes, we designed our library by targeting CRISPRi complex to open reading frame (ORF) rather than more popular TSS, thus to better investigate the responses of multiple genes in an operon. Following selection under two different conditions (minimal and acidic medium), we identified most of the known gene-phenotype associations, suggesting that this platform provides both high specificity and sensitivity in a high-throughput manner. For operon structures in *E. coli* genome, our results showed that by checking the sgRNA abundance profiles targeting the ORF region of every gene, rather than the common promoter of the particular operon, it is possible to narrow down the number of candidate genes contributing to the phenotypic effect. Moreover, we extracted several new rules for sgRNA library design for such pooled functional genomics screen by exploring the comprehensive sgRNA activity dataset produced from the screens. One

interesting observation was that only sgRNAs targeting the first 5% of ORF showed enhanced repression activity. Additionally, we established that ten sgRNAs per gene were enough for robust hit-gene identification by computational subsampling. Finally, we optimized the hit-gene calling algorithm by only recruiting sgRNAs with better activities. This work established that CRISPRi based pooled screen can be a powerful platform for high-throughput functional genomics study in prokaryotes.

Results

Design of sgRNA library and CRISPRi screen system

To test whether CRISPRi-based pooled screen can be applied to *E. coli*, we wanted to design an sgRNA library targeting genes for which a knockout produces a known and easily selectable phenotype, such as a change in growth under specific cultivation conditions. To this end, we turned to the Keio library(29), an exhaustive *E. coli* gene knockout collection with extensive phenotypic characterization. Gene knockouts with significant impacts on growth in acidic medium were extracted from the dataset reported by Nichols et al(30), and auxotrophic gene knockouts in MOPS medium (OD_{600} < 0.1 after growth for 24 and 48 h in MOPS medium) were identified from the original paper describing the Keio library(29). All genes thus

147 identified were cross-checked to verify normal growth in LB broth(29). To determine whether
148 dense coverage of sgRNAs in multiple genes belonging to one operon can be used to assess
149 genotype-phenotype associations at higher resolution, we used the RegulonDB database(31) to
150 check the operon structure of the selected genes.
151
152 Using these lines, we selected genes transcribed as monocistronic mRNAs with impaired
153 growth in MOPS medium to generate Library I, which consisted of 22 candidates. We also
154 selected a series of genes residing in operons transcribed as polycistronic mRNAs carrying
155 auxotrophy in MOPS medium, as well as all their co-transcribed partner genes without relevant
156 phenotypes, to generate Library II, which consisted of 22 genes from nine operons with one
157 auxotrophic gene per operon. Furthermore, we selected genes with impaired growth in acidic
158 medium (indicated by a change in colony size on agar plates)(30) to produce Library III, which
159 consisted of 23 genes. Genes in all three libraries are listed in Table S1. Libraries I and II were
160 used as a proof-of-concept for CRISPRi-based pooled screen for lethal phenotypes. In contrast,
161 Library III was used to show the power of pooled screen to confirm moderate
162 phenotype-associated genes with either positive or negative growth effects.
163
164 A customized python script was developed to design up to 50 sgRNAs for each gene from the
165 start codon in the open reading frame (ORF) targeting the non-template strand, due to the better

166 activity than template strand as reported previously(16), generating sgRNA Libraries I, II and
167 III. A negative control sgRNA library (Library NC) was also designed, consisting of 400
168 N20NG(A)G 23mers with at least five mismatches from any site in the *E. coli* genome. All
169 designed sgRNAs are listed in Table S2, and basic statistics for the sgRNA library are
170 presented in Table 1. The sgRNA library was synthesized by DNA microarray and
171 incorporated into the pTargetF_lac expression vector by Golden Gate Assembly(32).
172
173 For the CRISPRi system, we used a constitutively expressed dCas9 protein (J23111 promoter)
174 and the leaky expression of sgRNA from a $P_{L\text{-}lacO}$ promoter (Figure S1a) after several rounds of
175 optimization (see Methods). The activity of this system was then tested in diverse constructs
176 targeting *sfGFP* (integrated at the *smf* locus) (Figure S1b), *crtE* (integrated at the *ldhA* locus) in
177 the lycopene biosynthesis pathway (Figure S1c) and *sacB* (integrated at the *smf* locus) with
178 cellular toxicity in the presence of sucrose (Figure S1d). Our results confirmed that this system
179 could be applied to repress gene expression from diverse loci in the *E. coli* chromosome. To
180 confirm the impact of gene knockdown on cell growth, we individually constructed a series of
181 sgRNAs from Library I or II, transformed the plasmids into *E. coli* MCm carrying
182 pdCas9-J23111, and measured growth curves in selective medium (MOPS, Figure S2). Our
183 results show that the majority of sgRNAs, together with dCas9 protein, significantly impaired
184 growth compared to the negative control sgRNAs, indicating that CRISPRi-derived repression

can provide sustained phenotypic effects to be detected by pooled screen in a high-throughput manner.

With sgRNA libraries and the CRISPRi system in hand, pooled screen for high-throughput functional genomics in *E. coli* was performed, as shown in Figure 1. *E. coli* MCm, a K12 MG1655 derivative with a chloramphenicol-resistance cassette integrated (see Methods) was chosen as the host cell in this work, considering that the chloramphenicol marker can be used to detect contamination. The sgRNA library plasmids were prepared and transformed by electroporation into *E. coli* MCm carrying pdCas9-J23111, giving rise to four cell libraries (designated I, II, III and NC, according to the relevant sgRNA library). The cell libraries were then mixed and subjected to relevant selection conditions. The change (selective vs. control conditions) in relative abundance of each sgRNA (sgRNA fitness) was resolved via deep sequencing. Based on the information of sgRNA fitness belonging to each individual gene, the quantitative estimation of genotype-phenotype association (median sgRNA fitness) was calculated and the statistical significance was determined by comparison with negative control sgRNAs.

Identification of genes with expected auxotrophic phenotypes from pooled screen

Cell libraries I, II and NC were combined for selection in minimal medium (MOPS) and constituted the 'Minimal Library'. Two biological replicates were included for each selection condition. A pooled screen approach (Figure S3) was applied, with approximately ten cell doublings (OD_{600} from 0.001 to ~1.0; 6.5 h for control while 24.2 h for selective condition), in principle enabling two-fold enrichment or dilution for mutants with as little as 7% fitness change per generation. After selection, plasmids were extracted, sequencing libraries were constructed and deep sequencing was carried out. For all libraries, ~80% of reads were mapped back to the synthetic sgRNA library (Figure S4), suggesting that the sequencing procedure was sufficiently reliable. Obvious selection pressure was observed for libraries under selective conditions, as determined by evaluating library Gini indexes before and after selection (Figure S4). The consistency of data from two biological replicates was also confirmed in terms of normalized sgRNA read number (Figure S5).

We used the relative change in abundance of each sgRNA (*rho* score, a quantitative estimation for sgRNA fitness, proportional to the relative change of sgRNA abundance; see Methods) between selective and control conditions to deduce the impact of sgRNA on growth via repressed gene expression. All processing data (*rho* and Z scores for each sgRNA) for Minimal Library selection is reported in Table S3. The *rho* score profile for sgRNAs targeting genes with auxotrophic phenotypes in MOPS medium diverged significantly from the profile for the

223 negative control set (Figure 2a, Figure S6). To determine statistical significance, we applied
224 the Mann-Whitney U test(33) for comparison between *rho* score distributions of sgRNAs
225 targeting a given gene and the control sgRNA set, leading to *P* value evaluations of the
226 phenotypic impact of each gene (Figure 2b, Figure S6). In addition, to perform multiple testing
227 corrections, 'quasi genes' were constructed from the control sgRNA set, and the false
228 discovery rate (FDR) was measured for different *P* value thresholds as well as a variety of
229 sgRNA numbers (Figure S7). This analysis showed that sgRNA number does not significantly
230 impact the profile of FDR curves, and that FDR reached ~1% with a *P* value threshold of 0.01.
231 Hence, a *P* value of 0.01 was used as the threshold to call statistically significant
232 genotype-phenotype relationships in the subsequent work, unless otherwise indicated. Using
233 this threshold, we found that the majority (16/22) of genes from Library I that are auxotrophic
234 in MOPS medium could be recovered (Figure 2b). Meanwhile, only one gene unrelated to
235 auxotrophy in Library II had a significant *P* value (Figure 2b), giving a false positive rate (1/13)
236 that was similar to the FDR determined by multiple testing correction (~1% with 0.01 *P* value
237 threshold). These results demonstrate that our approach is not only highly sensitive, but also
238 highly specific for high-throughput genotype-phenotype association.
239
240 We next sought to determine the mechanism behind several false negative results in screen.
241 Intriguingly, among false negative genes in Library I, genes in the arginine biosynthesis

242 pathway (*argA*, *argE* and *argG*) were significantly enriched ($P = 0.0065$ by one-sided Fisher's
243 exact test). However, pure culture growth tests showed that representative sgRNAs targeting
244 these three genes, together with dCas9 protein, significantly blocked cell growth (Figure S2). A
245 search of the EcoCyc database revealed that ten arginine transporters are encoded in the *E. coli*
246 genome, suggesting that strains with normal arginine metabolism rescued the growth of
247 arginine auxotrophic mutants by intermediate cross-feeding in the consortium during screen.
248 To test this hypothesis, we cultivated wild-type *E. coli* (pdCas9-J23111+ control sgRNA
249 plasmid) in MOPS medium for 24 h and recovered the supernatant medium by centrifugation
250 and filter-sterilization. We verified that there was no growth in this sterilized medium without a
251 seed culture. Then, we mixed this sterilized medium with fresh MOPS medium in a series of
252 proportions and used this for growth tests (Figure S8). We found that a 1:1 (v/v) mixture
253 enabled growth of a series of gene-knockdown strains, including those for *argA*, *argE* and
254 another false negative gene identified in the screen, *lysA*. It is noteworthy that this syntrophic
255 exchange experiment only partly mimicked the pooled screens. Differences did exist such as
256 the availability amount and timing of key exchanging intermediate(s) between these two
257 experimental settings, which could result in some inconsistent results such as the recovered
258 growth of true positive *cysG* identified in pooled screen and failure to restore the growth of
259 *argG* knockdown (false negative in pooled screen). Even though, this result strongly suggests
260 that cross-feeding masks the growth deficit of auxotrophic mutants in the consortium, a

reasonable but commonly underestimated pitfall of pooled screens, although further investigation is needed to determine the specific contributing metabolite(s). In this sense, these genes are not true false negatives because inherently they cannot be identified as hit for the phenotype in a pooled screen format. As a supporting evidence, a paper reporting a synthetic *E. coli* co-culture system also indicated such possibility by observing similar syntrophic exchange phenomena between arginine auxotrophs and several other amino acid auxotrophic mutants(34).

Gene-phenotype association in operons with polycistronic mRNA transcribed

We hypothesized that by checking sgRNA fitness profiles for the genes within the same polycistronic-mRNA-transcribing operon using data from pooled screen, it is possible to narrow down the candidate gene pool responsible to the studied phenotype (auxotrophy in MOPS medium in this case). For example, according to the current CRISPRi working model as a transcriptional roadblock indicated by RNA polymerase nascent RNA sequencing results(16), if the functionally responsible gene locates at the 3' downstream or middle of the operon, all other genes upstream of this one in the operon should exhibit perturbed sgRNA fitness profile (Figure S9, upper and middle panel), because the repression of these genes by CRISPRi also cause knockdown of downstream target gene. In contrast, if the functionally responsible gene

locates at the upstream of the operon, only this gene is expected to be identified as hit while all others downstream not (Figure S9, lower panel).

To test this hypothesis, screen data from Library II was assessed. There are nine operons with polycistronic mRNA transcribed in Library II (Table S1). Among five of them (*nirBDC_cysG*, *glnALG*, *serB_radA_nadR*, *nadA_pnuC*, *aroF_tyrA*), genes with known growth impairment phenotypes were identified as hits (gave $P < 0.01$, Mann-Whitney U test) with negative median sgRNA *rho* scores, suggesting successful functional association in these cases (sensitivity = 5/9, Figure 2b, 3). Including these five operons, *serB_radA_nadR* and *nadA_pnuC*, where the auxotrophic genes both reside at the upstream of the particular operon, followed the expected sgRNA fitness profile (as elucidated in Figure S9, lower panel). Among the three remaining operons, in spite of successful functional gene mapping, sgRNA fitness profiles partly different from our hypothesis were observed. For *nirBDC_cysG* and *aroF_tyrA*, due to the location of auxotrophic genes at the 3' downstream of the operon, it is expected that all genes in the relevant operons should carry perturbed sgRNA fitness profile (Figure S9, upper panel). However, we surprisingly found that only *cysG* and *tyrA* exhibited reduced sgRNA abundance. One possible reason for this conflict with known CRISPRi working mechanism is that activity from unknown promoters drives downstream gene (*cysG* and *tyrA* in this case) expression, which would not be repressed by *trans*-elements regulating upstream gene expression. For

299 instance, *cysG* has two known promoters right upstream of its coding region (cysGp1, cysGp2,
300 RegulonDB). For *aroF_tyrA* operon, diverse transcript profiles have been described in this
301 operon. RNA-seq data in RegulonDB (but not determined experimentally) suggested the
302 existence of TSS_2941 promoter, residing in the *tryA* ORF region of the *aroF-tyrA* operon.
303 However, another recent RNA-seq experiment only identified the known aroFp promoter(35).
304 Additionally, in *glnALG* operon, *glnL*, in spite of moderate statistical significance and
305 unexpected positive median sgRNA *rho* score, was identified as hit in pooled screen. However,
306 according to the pattern of gene location in this operon (Figure S9, lower panel), only *glnA*, the
307 gene functionally related to auxotroph at the upstream of the operon, should be identified as hit.
308 The molecular mechanism for this phenomena might be due to the reverse polar effect
309 suggested in a previous work(23), where the repression of downstream *gfp* in an artificial
310 *rfp-gfp* operon by CRISPRi unexpectedly gave rise to the perturbation of upstream *rfp*
311 expression. The relatively weak property of reverse polar effect(23) is also consistent with our
312 observation that *glnL* only exhibited moderate association with auxotrophy.
313
314 For the other four operons (*fic_pabA*, *pyrBI*, *rsmE_gshB*, *kbaZ_agaV*), no genes showed
315 significant changes in sgRNA *rho* score profiles, except for *pabA* with a positive median
316 sgRNA *rho* score but only minor statistical significance (thus a false positive as defined in
317 Methods). The expected phenotype-determining genes include *pabA*, *pyrB*, *rsmE* and *agaV*.

One reason for the false negative result of *pabA* and *pyrB* might be poor CRISPRi-based repression. We reconstructed two sgRNAs targeting these two genes and determined the growth profile in MOPS medium. We failed to observe any difference from the control group (Figure S10). For *rsmE* and *agaV*, knockouts result in only moderately impaired fitness in MOPS medium(29). Although the *rsmE* knockout mutant is defective compared to the wild-type strain(36), such a moderate difference in fitness might only be detectable with prolonged selection time (currently, ten doublings) or more stringent selection pressure. It is very interesting to observe a much higher false negative rate (4/9) for operons with polycistronic mRNA, in contrast to coping with operons with single gene (6/22). This result suggested the complexity of coordinated transcription of polycistronic-mRNA operons and unknown interaction of such complexity with CRISPRi function.

CRISPRi screen for robust identification of moderate-phenotype-associated genes

To explore the capacity of CRISPRi-based pooled screen to map genes to moderate phenotypic effects, we combined cell Libraries III and NC as a 'LowpH Library' and subjected it to selection in acidic medium (LB4.5 medium, see Methods). We applied a pooled screen approach (Figure S3) with approximately five cell doublings (OD_{600} from 0.01 to ~0.4; 2.3 h for control while 9.0 h for selective conditions), enabling two-fold enrichment or dilution for

mutants with as little as 15% fitness change per generation. All processing data (*rho* and Z scores of each sgRNA) for LowpH Library selection is reported in Table S4. As in the Minimal Library screen, the mapping ratio, selection pressure (Figure S11) and biological replicate consistency (Figure S12) were acceptable. We also evaluated the FDR curve by constructing quasi-genes from the negative control sgRNA set (Figure S13). We applied a 5% FDR (fit by interpolation from data in Figure S13) as the threshold for hit-gene calling and found that a significant fraction (9/23) of known growth-perturbing mutants in acidic medium can be recovered by pooled screen (Figure 4). The false negative results observed might be due to the systematic difference between physiological responses to acidic stress in liquid medium (this work) and on agar plates (dataset used to extract genes in library III)(30). This result shows that our method is a powerful tool for constructing gene-fitness maps when gene expression perturbation has only a moderate impact on growth (the majority of median *rho* score absolute values were less than 0.02, whereas those of more than half of the true positives in the Minimal Library were higher than 0.05; Figure 2b).

Overview of sgRNA activity landscape across ORF

With the dataset produced in abovementioned screen, we sought to further address the sgRNA activity issue, because we observed in our dataset great sgRNA activity diversity (Figure 2a),

as reported in previous work(cite). We firstly checked the effect of sgRNA location within ORF, an important feature found to determine sgRNA activity in CRISPRi system(16) but only assessed by case study rather than big data thus far. To this end, we combined sgRNAs from Library I whose corresponding genes are shown to be true positives, thus constructing a 'functional' sgRNA set (16/22 genes, 468 sgRNAs; Figure 2b). The absolute values of sgRNA Z scores (see Methods) are a reasonable metric to evaluate their activities. We categorized sgRNAs in this set into subgroups according to their relative position along the ORF. We then examined the difference in activity between each subgroup and the whole population using the Mann-Whitney U test (Figure 5). We observed that only the sgRNA subgroup located within the first 5% of the ORF region exhibited enhanced activity ($P = 0.0030$, threshold $P < 0.01$), whereas all other subgroups did not. This was consistent with previous reports indicating that sgRNAs targeting upstream regions of the ORFs exhibited better activity(16). Our results, which are based on comprehensive big-data analyses, define this optimal window for active sgRNA positioning with better resolution compared with previous works. It should be noted that this dataset is highly noisy due to the functional consequences of gene knockdown are inherently diverse across genes. Considering the importance to select highly active sgRNAs incorporated into the library, we suggested the need to develop more unbiased strategy to differentiate the knockdown activity of sgRNAs(37), enabling better design of synthetic sgRNA libraries.

375

376 **Number of sgRNAs per gene needed for robust gene calling**

377

378 Reducing the number of sgRNAs per gene is expected to reduce the cost of library preparation
379 and facilitate handling in large-scale experiments. In order to determine the minimal sgRNA
380 set needed for reliable hit-gene calling, we performed a subsampling approach based on
381 sgRNAs targeting the true positive genes in Library I (16/22 genes, 545 sgRNAs) collected
382 from Minimal Library screen. Five strategies were tested to determine the priority of sgRNA
383 selection during subsampling: 'Position', 'Random', 'RS2', 'Cas9cal' and 'SSC'. The Position
384 strategy chooses the sgRNA set most proximal to the start codon of the ORF, where more
385 activity can be expected (Figure 5). The Random method selects sgRNAs randomly during
386 subsampling. Cas9cal(38), RS2(27) and SSC(39) determine sgRNA subsampling priority
387 based on scores given by a previously reported sequence-activity machine learning model. For
388 SSC, we chose a model specially trained from eukaryotic CRISPRi data. For Cas9cal, we used
389 a hybrid model of dCas9 binding and Cas9 cleavage activity. In contrast, RS2 was trained
390 exclusively from CRISPR/Cas9-induced DNA cleavage–based loss-of-function screen data.
391 We made a hypothesis here that sgRNA activities in CRISPR/Cas9 system, to a certain extent,
392 positively related to the activities in CRISPRi system.

393

We subsampled sgRNAs for each gene based on these five strategies and then calculated the *P* value in comparison to the 400-member control sgRNA set by the Mann-Whitney U test (Figure 6). We found that the Position method generally outperformed all others, especially when only a minimal sgRNA set was available. For instance, when 5 sgRNAs were available, except for Position method, all other subsampling approaches resulted in a big portion of genes identified as false negative due to the *P* value below the threshold (Figure 6, below the dashed line). The differences among the five methods became smaller when more sgRNAs were sampled. When the sgRNA subset reached ten members, nearly all genes reached the significance threshold using the Position subsampling strategy ($P < 0.01$ for 14/16 genes, giving FDR ~ 0.01, as shown in Figure S7; $P < 0.05$ for remaining two genes). This suggests that ten sgRNAs per gene is sufficient for robust functional genomics screen, at least under our selection scenario.

Optimization of gene-calling algorithm

Contrary to our expectation, during Position-based sgRNA subsampling, we found that a subset of sgRNAs occasionally resulted in stronger statistical significance than all available sgRNAs. Figure S14 presents the subsampling profiles for four representative genes. For *purC* and *purM,* the main sgRNA contributing to gene detection resides at the 5′ region of the ORF,

413 suggesting that a more complete knockdown is necessary to impair cell growth. In contrast, for
414 *gltA* (a strong hit-gene encoding citrate synthase) and *cysB* (a weak hit-gene regulating cysteine
415 biosynthesis and sulfur metabolism), the statistical significance generally increased as more
416 sgRNAs were tested. This phenomenon indicates that even partial knockdown of gene
417 expression lead to an observed phenotype in these cases. This observation reveals that the
418 efficiency of CRISPRi-based screen relies on multiple factors. When a significant change in
419 gene activity is needed to cause a phenotype, sgRNA activity is the dominant contributing
420 factor in statistical significance. In contrast, when moderate perturbation of gene expression is
421 sufficient, sgRNA number is the dominant factor.
422
423 Based on these observations, we hypothesized that a properly selected subset of sgRNAs
424 belonging to a gene might give stronger statistical significance than the intact set. Hence, we
425 tried to optimize the hit-gene calling algorithm by applying the Position strategy to subsample
426 the sgRNA set and searching for the peak of statistical significance (by Mann-Whitney U test)
427 with at least five sgRNAs to ensure robustness (Figure S15). We compared the performance of
428 this novel strategy with the previous version as a reference for hit-gene calling in the Minimal
429 Library screen dataset (Figure S16). We used only the first 15 sgRNAs at the 5′ region of the
430 ORF to assess the capacity of the new algorithm to identify hit genes with fewer sgRNAs. The
431 optimized approach maintained the performance of hit-gene calling (compared with Figure 2b)

with fewer sgRNAs, and improved the statistical significance of hit-gene calling for a subset of genes (*purC*, *purL*, *purM* and *nadA*), all of which have a typical sgRNA subsampling profile with a statistical significance peak between 5 and 15 sgRNAs. This optimized strategy also maintained performance in determining genotype-phenotype associations in complex co-transcriptional units (Figure S17 vs. Figure 3 produced by previous gene-calling algorithm with all available sgRNAs). To evaluate the robustness of the new algorithm applied to moderate phenotype mapping, we repeated the pipeline using only the first 15 sgRNAs most proximal to the start codon of relevant genes in Library III. We applied a 5% FDR threshold (fit by interpolation from data in Figure S13) and identified eight positive hit-genes among all 23 candidates in this dataset (Figure S18), six of which were originally identified as hits by the previous program with all available sgRNAs (Figure 4). This demonstrates both the reliability and improvement (two additional hits with a smaller sgRNA set) of the optimized algorithm.

Discussions

This work, for the first time in prokaryotic organisms, presents a novel high-throughput functional genomics platform enabled by CRISPRi pooled screen. Our data supports that this method is both highly sensitive (16/22 coping with genes transcribed as momocistronic mRNA, while 9/23 for genes with moderate phenotypic effect as demonstrated in acidic

451 medium selection) and specific (at least < 5% FDR). Importantly, our library design considers
452 the unique genome structure of prokaryotic organisms such as that we target sgRNAs to ORF
453 rather than regions around TSS, which is widely adopted in eukaryotic sgRNA library design
454 for CRISPRi screen. This feature, at functional level, enables us to investigate the sgRNA
455 fitness profile of each individual gene in polycistronic-mRNA transcribing operons, thus
456 narrowing down the candidate gene pool for the studied phenotype (successful in 5 of 9 cases).
457
458 CRISPRi technology provides new opportunities and advantages for high-throughput
459 functional genomics in prokaryotic cells, compared with several other established pooled
460 screen methods based on transposon gene knockout(1, 5) or recombineering(3), in terms of
461 better applicability to more microorganisms, customized library design, easier preparation and
462 handling of library as well as less bias problem. Indeed, highly efficient recombineering toolkit
463 is only available to several model bacterial strains with either clinical or industrial importance
464 after decades of mining and optimization(40–43), since its introduction nearly 20 years
465 ago(41). In contrast, only a working expression system is needed for CRISPRi based functional
466 genomics platform to be established in a new prokaryotic organism, and this system has been
467 shown to be broadly applicable from prokaryotic(7, 20–23) to eukaryotic(8–11) cells in the
468 recent four years. In the aspect of bias issue, a benchmark work constructing a comprehensive
469 transposon mutant library in *Burkholderia thailandensis* presented 86.7% coverage for 5,634

predicted genes after optimizations(44), given that 7.1% of all genes as putatively essential(45). In our case, we achieved > 99.5% coverage as well as 88.1% 10-fold variation in the prepared library for the designed sgRNA set in this work. Moreover, these parameters were maintained for a genome-wide sgRNA library targeting > 4,000 genes in *E. coli* without dCas9 expression (unpublished data). This highly even distribution of mutant abundances and the customized library design make it possible to study the functions of focused gene set with short coding length, such as (a)sRNAs, which have been shown to play important roles in the responses to environmental changes in prokaryotic organisms(46, 47). In fact, strategies applying CRISPR/(d)Cas9 method for high-throughput non-coding RNA functional profiling have been recently reported in the human genome screen(48, 49). During the preparation of our manuscript, a report described a genetic mapping analysis via pooled screen and deep sequencing based on CRISPR/Cas9–facilitated recombineering in *E. coli* (CREATE)(50). Although CREATE is a very powerful platform, it suffers from inherent variability in recombination efficiency across genomes (like the TRMR approach(6)), which can cause problems in downstream data analysis. Hence, our method provides the first proof-of-concept of CRISPR/(d)Cas9 based high-throughput functional genomics in prokaryotic genetics at gene level, in contrast to CREATE focusing at nucleotide level. Except for the advantages presented in this work, our proof-of-concept method still suffers from shortcomings such as inability to induce gene overexpression or expression level modulation, which is important to

more comprehensively profile a genome as shown in other platforms(6, 51). However, the versatility of CRISPR/(d)Cas9 based system have been extensively demonstrated(17, 52). Further works need to be carried out to realize these potentials for more powerful functional genomic screen in prokaryotes.

This work presents the first guideline learned from the big-data for the construction of sgRNA libraries for CRISPRi pooled screen towards identifying genes important for particular phenotypes in prokaryotes. For example, we successfully identified the optimal window for highly active sgRNAs at the first 5% of ORF region. Meanwhile, the low false positive rate of our screen result suggests that sgRNA off-target effect based on current cutoff setting is minimal, especially considering the fact that typical prokaryotic genome is much smaller than that of mammalian cells. We also suggested 10 sgRNAs per gene in library design by computational subsampling analysis. Currently, most of the sgRNA design tools focus specially on eukaryotic organism, such as human or mouse cells(27, 53). Fewer available tools addressing the usage of CRISPR system in prokaryotes also basically depend on the rules for eukaryotic organisms. For example, in CRISPR-ERA(54), a popular webserver able to provide batch sgRNA design for genome editing and repression for *E. coli* and *B.subtilis*, the only rule regarding sgRNA efficacy is their position within a optimal window 500bp downstream of the TSS, which does not consider important features for CRISPRi usage in prokaryotes such as

operon structure (targeting around TSS generally represses the transcription of the operon), strand (targeting non template strand was shown to have higher CRISPRi activity(16)) and the position-activity relation for sgRNA activity (Figure 5, in another aspect, for prokaryotic gene without any intron, 500bp sometimes covers the whole gene-coding region). This work presented the preliminary solutions of these issues via the abovementioned library design guideline. We are also working to produce bigger dataset to train models and construct user-friendly software packages, aiming at providing with the community currently unavailable tools to design better sgRNA (libraries) in prokaryotes.

For polycistronic-mRNA-transcribing operons, a very common organization of prokaryotic genes in genome ,our approach successfully narrowed down the target gene pool. However, we also observed exceptions of sgRNA fitness profile patterns at functional level, such as the absence of gene repression when targeting CRISPRi complex to the upstream genes (*nirBDC_cysG* and *aroF_tyrA*), a potential conflict with known CRISPRi mechanism. One potential explanation for this phenomenon is the unidentified promoters within the operon, like cysGp1 or 2. Moreover, some recent papers suggested that dCas9-sgRNA complex could bind ssRNA, in spite of much lower affinity(55–57). Due to our library design to target the non-template strand, the complex could theoretically bind the corresponding mRNA, thus independently inhibiting the translation process of each individual coding region within the

polycistronic mRNA, which might be an alternative reason for the unexpected sgRNA fitness profile pattern mentioned above. Because of the general existence of operon structure and their importance at transcription regulation in prokaryotic genomes, further investigation is needed to more comprehensively understand the potential interactions when applying CRISPRi technology to operon structures at molecular level. We recommend that CRISPRi-based arrayed or pooled screen can be used to study the gene expression regulation in operons, as a complementary method to gene knockout (DNA sequence) and antisense RNA technology (currently known transcription vs. mRNA stability and protein translation)(58) functioning at a different level, the combination of which should give us novel insight into the mechanism of expression regulation in operons, such as the unidentified promoter activities within the operon suggested by our screen result (Figure 3, *nirBDC_cysG* and *aroF_tyrA* operons).

Another interesting observation is that the performance of the three machine learning models (Cas9cal, RS2 and SSC) we adopted for sgRNA subsampling is not significantly different from that of random strategy (Figure 6), and even exhibits poorer performance when less than ten sgRNAs are available for each gene. Based on current data we cannot rule out the contribution of the inherent difference in sgRNA activities when regarding CRISPR/Cas9 based DNA cleavage and CRISPRi based gene repression. However, considering the fact that SSC model was trained from purely eukaryotic CRISPRi data and the comprehensively reviewed structural

differences of eukaryotic and prokaryotic genomes(24), the failure of current models trained exclusively based on screen data from mammalian cells might be derived from the systematic difference between *in vivo* sgRNA activities on eukaryotic versus prokaryotic chromosomes. Indeed, even for eukaryotic cells such as yeast, a recent work identified a different optimal window relative to TSS for active sgRNA positioning in CRISPRi system compared with reported in human cell lines(59). Similarly, a recent report also stated the poor prediction of target activities (CRISPR/Cas9 derived cleavage) in *E. coli* by models trained from eukaryotic cell screen data(60). Hence, CRISPR/Cas9 or CRISPRi screens need to be performed in prokaryotic organisms as what have been extensively tested in eukaryotic cells. Such efforts will not only shed light on the sequence-activity relationship of sgRNAs in this poorly explored system to guide more efficient genetic engineering and function exploration, but will also more comprehensively elucidate the working mechanisms of the CRISPR system.

In summary, we designed, synthesized and screened an sgRNA library based on CRISPRi technology in *E. coli*, representing the first CRISPR/dCas9–based pooled screen functional genomics study in a prokaryote. Our results reveal that the synthetic sgRNA library–based pooled screen provides a powerful tool to facilitate high-throughput microbial genetic studies (Figure 2, 3 and 4). Moreover, our dataset, based on big-data analysis, revealed several novel insights for the design of sgRNA libraries for CRISPRi utilization in prokaryotic organisms,

565 such as activity-position relationships (Figure 5) and the poor performance of current activity
566 prediction models upon our dataset (Figure 6). The lessons learned from this work—including
567 establishing a threshold of ten sgRNAs per gene for robust hit-gene calling (Figure 6) and
568 development of an optimized hit-gene calling algorithm (Figure S15~18)—should facilitate
569 genome-wide sgRNA library design and data parsing for high-throughput functional genomics
570 studies (using gene knockdown or overexpression(61)) in *E. coli* and other important
571 prokaryotes.

572

573 **Materials and Methods**

574

575 **Strains, DNA manipulations and reagents**

576

577 *E. coli* MG1655 (wild type) was obtained from the ATCC (700926). *E. coli* s17-1 sfGFP was a
578 kind gift of the George Guoqiang Chen laboratory at Tsinghua University(19). All cloning
579 procedures were performed according to the manufacturers' instructions. DNA purification
580 (D2500) and isolation of high-quality plasmids (D6943) were performed using reagents from
581 Omega Bio-Tek (U.S.). DNA restriction and amplification enzymes were from New England
582 Biolabs. During plasmid construction, *E. coli* DH5a (BioMed) served as the host and was
583 cultured in Luria-Bertani (LB) broth or on LB-containing agar plates at 37°C. Plasmids were

584 constructed by Gibson assembly with a customized recipe, as described(62). Antibiotic
585 concentrations for kanamycin, ampicillin and chloramphenicol were 50, 100 and 7 mg/L,
586 respectively. MOPS medium was prepared as described(63). LB4.5 medium was prepared by
587 supplementing LB broth with 0.1 mol/L 4-morpholineethanesulfonic acid (Sigma M2933),
588 adjusting pH to 4.5 using hydrochloric acid, followed by filter sterilization. All cultivation was
589 carried out at 37°C.
590
591 **Strain and plasmid construction**
592
593 All strains, plasmids and primers are listed in Tables S5 and S6. *E. coli* strain MCm used in
594 library screen was constructed by inserting a chloramphenicol expression cassette cloned from
595 pKM154(64) (Addgene plasmid #13036) into the *smf* locus of wild-type *E. coli* K12 MG1655
596 by λ/RED recombineering(41). *E. coli* Msac was constructed by inserting a *sacB* expression
597 cassette (in which the J23105 promoter drives expression of *sacB* cloned from pKM154(64),
598 Addgene plasmid #13036) by CRISPR/Cas9 recombineering(65). *E. coli* lyc001 is a
599 lycopene-overproducing strain created by integrating a heterologously overexpressing *crtEIB*
600 cluster (cloned from pTrc99a-*crt*-M(66)) into the chromosome (unpublished data). The
601 nuclease activity–deficient Cas9 mutant dCas9 was derived from the pdCas9-bacteria vector
602 (Addgene plasmid #44249)(16). The promoter and resistance marker region were replaced

603 with a constitutive promoter (wild-type promoter for Cas9 from *Streptococcus pyogenes*) and
604 kanamycin marker cloned from pCas (Addgene plasmid # 62225)(65), resulting in
605 pdCas9-cons. The promoter was replaced by well-characterized iGEM Anderson promoters,
606 giving rise to plasmids pdCas9-J23(109, 111, 112, 113 and 116). The vector for sgRNA
607 expression was derived from pTargetF (Addgene plasmid #62226)(65) by replacing the
608 spectinomycin marker with an ampicillin expression cassette (pTrc99a(67)) lacking the *BsaI*
609 restriction site. The promoter region was substituted with a synthetic inducible promoter
610 ($P_{LlacO-1}$(68)) together with the corresponding repressor expression cassette *lacI* (pTrc99a(67)),
611 leading to pTargetF_lac. To facilitate library (amplified from oligonucleotides synthesized in
612 DNA microarray) insertion into pTargetF_lac, pTargetF_lac_preLib was constructed by
613 introducing two *BsaI* sites in opposite directions between the promoter and Cas9-binding site
614 region of pTargetF_lac. A series of pTargetF_lac plasmids targeting different genes was
615 obtained by inverse PCR with the modified N20 sequence hanging at the 5′ ends of primers
616 followed by self-ligation. To tune the induction profile of the CRISPRi system, we replaced the
617 native promoter upstream *lacI* with the strong constitutive J23100 promoter and the ribosome
618 binding site (RBS) with synthetic RBSs designed using RBS calculator(69) (RBS0-4)
619 presenting enhanced translation intensity (Figure S19, pdCas9-J23116 was used to reduce
620 dCas9 expression). The P_L promoter strength was also modulated by introducing mutations to
621 the –10 motif, as reported(70) (W for wild type, M for mutant).

622

623 **Optimization of CRISPRi system**

624

625 To reduce the noise introduced during selection, we aimed to developed a constitutive dCas9
626 expression plasmid, in contrast to currently reported dCas9 constructs under the control of
627 inducible promoters in prokaryotes(16, 19, 23, 71). At the same time, we tried to establish
628 inducible sgRNA expression systems to facilitate manipulation of essential genes (Figure S1a).
629 To this end, we constructed a series of dCas9 expression plasmids under the control of
630 constitutive iGEM Anderson promoters with a series of expression strength. We used strong
631 inducible $P_{L\text{-}lacO}$ promoters with a well-defined transcription start site and tight regulation to
632 drive sgRNA expression. However, we failed to observe any inducible knockdown activity
633 owing to expression leakage, and instead found that repression activity was generally
634 determined by the strength of dCas9 expression (Figure S1b, the repression level is
635 proportional to the strength of Anderson promoter, J23111 > J23116 > J23109 > J23113 >
636 J23112), in accordance with the assumption that this system being dCas9-limited and having
637 sgRNA in abundance. Further optimization of relevant repressor and P_L promoter expression
638 strength still failed to produce the expected induction property (Figure S19, pdCas9-J23116
639 was used to reduce dCas9 expression). These results suggested that a moderate level of sgRNA
640 expression was sufficient to drive sustained CRISPRi activity, which is consistent with the fact

641 that inducible CRISPR systems developed thus far in prokaryotes have been based on
642 regulation of dCas9 expression(16, 19, 23, 71) and the vector backbone we used here for
643 sgRNA expression has a relatively high copy number (pMB1 origin, ~15-20 copies/cell).
644 Based on the result of dCas9 constructs with diverse expression strength (Figure S1b), we used
645 the pdCas9-J23111 (the strongest promoter among the five constructs) and pTargetF_lac
646 plasmids for the following work, exploiting the leaky expression of sgRNA from the $P_{L\text{-}lacO}$
647 promoter, as this provides sustained repression activity, enabling further tuning, if required.
648
649 **Characterization of CRISPRi system**
650
651 For fluorescence characterization, overnight LB cultures (ampicillin and kanamycin) from a
652 single colony of *E. coli* s17-1 sfGFP containing relevant dCas9 and sgRNA (or control sgRNA
653 without potential target in *E. coli* genome) expression plasmids were individually incubated in
654 10 mL fresh LB medium in 50-mL flasks (initial OD_{600} = 0.02) with or without 1 mM isopropyl
655 β-D-1-thiogalactopyranoside. Subsequently, cells were cultivated for 12 and 26 h, and
656 fluorescence was measured with an F-2500 Hitachi Fluorescence Reader (excitation, 488 nm;
657 emission, 510 nm). Fluorescence was normalized to the culture OD_{600} value measured on an
658 Amersham Bioscience spectrophotometer. The repression ratio was calculated by comparison
659 of relative fluorescence to the control strain expressing the non-targeting sgRNA.

660

661 For lycopene accumulation characterization, overnight LB cultures (ampicillin and
662 kanamycin) from a single colony of *E. coli* lyc001 containing dCas9-J23111 and
663 pTargetF_lac_crtE1/2 (or control sgRNA without potential target in *E. coli* genome) were
664 individually incubated in 10 mL fresh LB medium in 50-mL flasks (initial OD_{600} = 0.02).
665 Subsequently, fermentation was carried out for 24 h and lycopene was measured as
666 reported(66). The titer was normalized to the culture OD_{600} value.

667

668 For growth testing of *E. coli* Msac, LB agar plates without sodium chloride were prepared by
669 adding 500 g/L filtering-sterilized sucrose stock solution to autoclaved sodium-chloride-free
670 LB broth (1.8% agar) until a final concentration of 100 g/L. A single colony of *E. coli* Msac
671 containing dCas9-J23111 and pTargetF_lac_sacB1/2 (or control sgRNA without potential
672 target in *E. coli* genome) was streaked and cultivated at 30°C for 24 h.

673

674 **Design, synthesis and processing of sgRNA library**

675

676 The genome sequence of NC_000913.3 was used for library design in *E. coli* K12 MG1655.
677 Sequences for 20-mers used as sgRNAs were designed by customized python scripts. The
678 SeqMap package(72) was used to check potential off-target sites of the designed sgRNAs by

679 searching for N20NG(A)G 23-mers in NC_000913.3 with a tolerance setting of five
680 mismatches. Customized scoring metrics inferred from previous reports(15, 73) and illustrated
681 in Figure S20 were designed to evaluate off-target sites identified by SeqMap. Briefly, the
682 protospacer region was divided into three regions (8, 5 and 7 nt, from 5' end to 3' end as
683 Region III, II, I, respectively) according to the distance to the protospacer-adjacent motif
684 (PAM). We set this scoring metrics because mismatches are generally better tolerated at the 5′
685 end of the 20-nt targeting region of the sgRNA than at the 3′ end (proximal to PAM)(74). If the
686 PAM site of the off-target 23-mer was 'NGG', the mismatch penalty in the abovementioned
687 three regions was set as 2.5, 4.5 and 8, respectively, and the setting for 'NAG' was 3, 7 and 10,
688 respectively. The off-target site was considered significant when Σ(penalty $\times$ mismatch) <
689 threshold (11 in this work); relevant sgRNAs were eliminated from further processing
690 accordingly. According to a recent report comprehensively assessing off-target effect of
691 CRISPRi system via a partially degenerate library of variants(15) where we adopted the
692 off-target threshold setting from, our metrics can ensure minimal off-target effect. We took a
693 more stringent off-target cutoff than this previous report(15). For instance, one mismatch in
694 Region I together with another in Region III were found to completely abolish off-target effect
695 of CRISPRi induced repression(15). Even such off-target is considered significant
696 corresponding to 2.5+8<11 in our off-target detection rules. The threshold to exclude 20-mers
697 based on GC content was set at <25% or >75%(15). For each gene, sgRNAs (N20NGG)

meeting the described principles were designed from 5′ upstream regions targeting the non-template strand of the ORF(16) until 50 sgRNAs were extracted (for a small fraction of genes, less than 50 sgRNAs were designed). The sgRNAs were named after 'gene_p' according to the position (p) of the first guanine within the PAM region (NGG) in the ORF (e.g., rsmE_9, N20 = GTTCAGGATGATAAATGCGG).

Based on these principles, we selected 22 genes transcribed as monocistronic mRNA with impaired growth phenotypes in MOPS medium and designed sgRNAs from them, giving rise to Library I. In addition, we selected a series of genes residing in polycistronic-mRNA-transcribing operons with impaired growth phenotypes in MOPS medium, as well as all their co-transcribed partner genes without relevant phenotypic effect, leading to Library II, which consisted of 22 genes from nine operons with one auxotrophic gene (MOPS medium) for each operon. Then, 23 genes from either mono- or polycistronic-mRNA-transcribing operons whose knockouts cause significant growth perturbation in acidic medium (indicated by a significant change in colony size on agar plates compared to control) were collected to produce Library III. Finally, 400 negative control sgRNAs (Library NC) were designed by subsampling random 23-mers with at least five mismatches to any site (N20N(AorG)G) in the *E. coli* genome with proper GC content (25–75%). This setting is based on an observation that five mismatches, even all located at the

717 Region III, were enough to completely abolish the sgRNA activity in CRISPRi system(15).
718 Considering the fact that more off-target sites have mismatches locate in seed (Region II and I)
719 or PAM region, thus exhibiting even lower activity and the relatively smaller E. coli genome
720 size (~4.6 Mbp) compared with mammalian cells, this cutoff setting is expected to minimize
721 the off-target effect.
722
723 It should be noted that there is systematical difference in considering off-target effect of
724 sgRNA design coping with CRISPR/Cas9 and CRISPRi. As off-targets are already very
725 unlikely to cause significant perturbation of other genes in CRISPRi, a relatively loose
726 off-target setting can be tolerated. However, potential off-target sites might introduce double
727 strand break, giving rise to lethality or gene knockout in CRISPR/Cas9 system. In this case,
728 more cautious treatment should be considered regarding off-target issue for sgRNA design. In
729 fact, when we use an alternative version of code described in this work to do at-home sgRNA
730 design for individual construct used in CRISPRi or CRISPR/Cas9 system, a more stringent
731 off-target threshold (20) is used, because sgRNA number is not that important in individual
732 sgRNA design compared with library design (more sgRNAs provide stronger statistical
733 power).
734

735 The designed sgRNAs were synthesized as oligomers on an Agilent microarray and
736 constructed as a plasmid library by Golden Gate Assembly(32) with *BsaI*-digested
737 pTargetF_lac_preLib as the backbone vector. The library plasmids were transformed by
738 electroporation into *E. coli* MCm carrying the pdCas9-J23111 plasmid. Briefly, *E. coli* MCm
739 cells containing pdCas9-J23111 were grown in LB at 37°C until an OD_{600} of 0.8 was reached.
740 The flask was then placed on ice and all subsequent steps were performed on ice. The cells
741 were collected by centrifugation, washed five times in ice-cold deionized water, and
742 resuspended in 15% glycerol to concentrate them 50-fold. Cells were divided into 400-μL
743 aliquots and transformed with 1 μg library plasmid using an Eppendorf 2510 Electroporator
744 with a pulse of 25 kV cm^{-1}. Electroporation was carried out three times for each of the four
745 sub-libraries. The transformed cells were allowed to recover for 1 h at 37°C, then streaked on
746 LB agar plates containing kanamycin, ampicillin and chloramphenicol. Plates were incubated
747 for 12 h at 37°C. The coverage for each sub-library is reported in Table S7. Colonies were
748 scraped from agar plates in LB medium with relevant antibiotics, washed and resuspended at 6
749 × 10^9 cells per milliliter. Aliquots of each library were mixed 1:1 (v/v) with 50% (v/v) glycerol,
750 frozen in liquid nitrogen and stored at −80°C.
751
752 **Screen and selection**
753

754 Freezer stocks were inoculated into 50 mL LB medium using 10^8 cells for each of the four
755 libraries. Cells were grown at 37°C to reach an OD_{600} of ~1, then collected by centrifugation
756 and washed with MOPS medium. Libraries I, II and NC were combined according to the
757 relative sgRNA number ratio to give the Minimal Library. The LowpH Library was prepared
758 similarly by combining Library III and NC. These two libraries were used for further selection,
759 and a fraction of each was mixed 1:1 (v/v) with 50% (v/v) glycerol and stored at −80°C as the
760 initial library until plasmid extraction. Separate cultures of 100 mL MOPS and LB medium
761 with kanamycin, ampicillin and chloramphenicol were incubated with 10^8 cells from the
762 Minimal Library (two biological replicates each). Similarly, LB and LB4.5 medium with
763 kanamycin, ampicillin and chloramphenicol were incubated with 10^9 cells from the LowpH
764 Library (also two biological replicates each). Cells were grown at 37°C to an OD_{600} of ~1 for
765 the Minimal Library and 0.4 for the LowpH Library. A 1-mL aliquot of each culture was
766 centrifuged, washed and resuspended. Cultures were mixed 1:1 (v/v) with 50% (v/v) glycerol,
767 frozen in liquid nitrogen and stored at −80°C.
768
769 **Deep sequencing and data processing**
770
771 Frozen cell stocks (Min_start, Min_MOPS/LB_R1/2, LowpH_start, LowpH_LB7/4.5_R1/2;
772 ten samples; Figure S3) were inoculated into fresh LB medium with kanamycin, ampicillin and

773 chloramphenicol and grown to an OD_{600} of ~0.8. The plasmids were extracted and subjected to
774 gel electrophoresis. The results confirmed the robust maintaining of both dCas9 and sgRNA
775 expression plasmids (Figure S21). The purified plasmids were used as a template for PCR to
776 amplify the N20 region of library sgRNAs (for Minimal Library: 50-μL reaction, 50 ng
777 template, PF/R_pTargetLacNGS_SE75, Q5 polymerase, NEB M0491L, 98°C 30 s, 20 cycles
778 [98°C 10 s, 52.4°C 30 s, 72°C 10 s], 72°C 1 min; for LowpH Library: 50-μL reaction, 50 ng
779 template, PF/R_pTargetLacNGS_SE50, Q5 polymerase, NEB M0491L, 98°C 30 s, 17 cycles
780 [98°C 10 s, 53°C 30 s, 72°C 10 s], 72°C 1 min). The sequencing library was prepared following
781 the standard protocol. High-throughput sequencing (Annoroad Genomics, Beijing, China) was
782 performed on an Illumina HiSeq 2500 by the single-end 75-bp (SE75) technique for the
783 Minimal Library, and on an Illumina NextSeq 500 by the SE50 technique for the LowpH
784 Library. Approximately 10 million reads were collected for each library, providing at least
785 1,000-fold coverage.
786
787 Customized python scripts were used to extract the 20-mer variable sequences and remove
788 those with mutations within upstream or downstream regions (2–4 bp each) from raw *.fastq*
789 files. MAGeCK-VISPR(75) was used for extracted N20 set mapping back to the designed
790 sgRNA library, general parameter calculation (Gini index and mapping ratio) and library
791 normalization. To increase statistical robustness, only sgRNAs with >100 reads in the initial

792 library were used for further analysis. Further sgRNA statistics and genotype-phenotype
793 associations were calculated based on the framework described by Kampmann et al.(33) with
794 home-made python scripts. Briefly, the averaged read numbers from the two biological
795 replicates were normalized after MAGeCK-VISPR processing. The *rho* score for each sgRNA
796 was calculated as presented by the four equations below, suggested by previous work(33).
797

$$N_{NC}^{control}(t) = N_{NC}^{control}(t_0) \times 2^{gt_{control}} \quad (I)$$

798

$$N_{NC}^{selective}(t) = N_{NC}^{selective}(t_0) \times 2^{(g-k)t_{selective}} \quad (II)$$

800

$$rho'_{sgRNA} = Log_2\left(\frac{abundance_{selective}}{abundance_{control}}\right)/kt_{selective} \quad (III)$$

802

$$rho_{sgRNA} = rho'_{sgRNA} - median\left(rho'_{NC\ sgRNA}\right) \quad (IV)$$

804

805 Firstly, we utilized recorded OD_{600} during screen in control condition, together with proportion
806 of all negative control sgRNAs parsed by sequencing in relevant condition to determine
807 parameter *g*, referring to growth rate of wild type cells in control condition (Equation I).
808 Subsequently, *k*, the reduction in growth rate of wild-type cells under selective conditions
809 compared with standard conditions was determined similarly using data in screen under

selective conditions (Equation II). Raw *rho* scores for all sgRNAs were calculated by determining the relative change of abundance between selective and control conditions, further normalized by *k*t* (Equation III). Finally, *rho* scores for all sgRNAs were further normalized by the median of raw negative control sgRNA (Equation IV).

Negative control sgRNA *rho* scores in each condition were fit by normal distribution, giving rise to the standard deviation (σ). The Z score for each sgRNA was calculated by normalization with the σ value for the relevant library. In the original version of the hit-gene calling algorithm, the *P* value for each gene was derived from a two-tailed Mann-Whitney U test of all sgRNAs targeting that gene against the non-targeting control set. In the case of Minimal Library screen, if the median *rho* score of all sgRNAs belonging to the relevant auxotrophy-related gene was less than 0 (because gene knockdown in this library led to auxotrophy in MOPS medium) and the Mann-Whitney U test resulted in a *P* value less than the threshold (0.01 *P* value), the gene was categorized as a 'true positive'. 'False positive' included genes with no auxotrophy by knockout but significant *P* values, or those 'true' auxotrophic genes with significant *P* values but positive median *rho* scores. 'False negative' was used to describe auxotrophic genes with insignificant *P* values. In the case of LowpH Library screen, a gene was considered significant when the FDR value of the Mann-Whitney U test was less than threshold (5% FDR). In the improved version of the algorithm, we subsampled the sgRNA set (adding sgRNAs one by one

to calculate *P* values) based on relative location within the ORF (from start to stop codon) and searched for the peak of statistical significance (Mann-Whitney U test *P* value), which was used as the final *P* value (or derived FDR). To ensure statistical robustness, at least five sgRNAs were used for each gene.

Overview of sgRNA activity landscape across ORF

With the dataset produced in abovementioned screen, we sought to further address the sgRNA activity issue, because we observed in our dataset great sgRNA activity diversity (Figure 2a), as reported in previous work(26). We firstly checked the effect of sgRNA location within ORF, an important feature found to determine sgRNA activity in CRISPRi system(16) but only assessed by case study rather than big data thus far. To this end, we combined sgRNAs from Library I whose corresponding genes are shown to be true positives, thus constructing a 'functional' sgRNA set (16/22 genes, 468 sgRNAs; Figure 2b). The absolute values of sgRNA Z scores (see Methods) are a reasonable metric to evaluate their activities. We categorized sgRNAs in this set into subgroups according to their relative position along the ORF. We then examined the difference in activity between each subgroup and the whole population using the Mann-Whitney U test (Figure 5). We observed that only the sgRNA subgroup located within the first 5% of the ORF region exhibited enhanced activity ($P = 0.0030$, threshold $P < 0.01$),

whereas all other subgroups did not. This was consistent with previous reports indicating that sgRNAs targeting upstream regions of the ORFs exhibited better activity(16). Our results, which are based on comprehensive big-data analyses, define this optimal window for active sgRNA positioning with better resolution compared with previous works. It should be noted that this dataset is highly noisy due to the functional consequences of gene knockdown are inherently diverse across genes. Considering the importance to select highly active sgRNAs incorporated into the library, we suggested the need to develop more unbiased strategy to differentiate the knockdown activity of sgRNAs(37), enabling better design of synthetic sgRNA libraries.

Subsampling of sgRNA set

We computationally subsampled sgRNAs from the available set for one gene using five priority principles. The Position method checked the location of the sgRNAs along the ORF and chose sgRNAs whose targets were most proximal to the start codon. For the Random strategy, ten subsamplings were carried out, and the average of the Mann-Whitney U test *P* values was calculated. For the other three approaches, the scripts for three sequence-activity machine-learning models (Cas9cal(38), RS2(27) and SSC(39)) were downloaded, and the

following commands were used to calculate an activity score for each sgRNA, allowing subsequent selection of the most 'active' candidates.

/bin/SSC -l 20 -m SSC0.1/matrix/human_CRISPRi_20bp.matrix -i input.txt -o output.txt

python rs2_score_calculator_v1.2.py --seq N4N20NGGN3

python Cas9_Calculator_batch.py crRNAseq PAM target quickmode=False, cModelName='All_dataModel.mat'

High-throughput growth curve measurements

All strains were grown overnight in deep 24-well plates containing LB medium with ampicillin, kanamycin and chloramphenicol. Cell pellets were collected by centrifugation, washed once with MOPS medium and resuspended. OD_{600} values were measured and cells were diluted in fresh MOPS medium with antibiotics at an initial OD_{600} of 0.01. Cells were grown with shaking at 37°C in a 96-well plate reader (Tecan 2500 Pro) and OD_{600} was measured at 15-min intervals.

Statistical information, software and figure generation

Plots were generated in Python 2.7 using the matplotlib plotting libraries and plotly online server (https://plot.ly/). Statistical analysis (two-tailed Mann-Whitney U test), data fitting and interpolation calculation were performed using the SciPy and NumPy Python packages.

Table and Figure Captions

Table 1 Basic statistics of the synthetic sgRNA library

	Library I	Library II	Library III	Library NC	Total
Number of genes	22	22	23	0	67
Number of sgRNAs	976	905	867	400	3148
Description	MOPS-auxotrophic	MOPS-auxotrophic-operon	Growth perturbation in acidic medium	Negative control	

Figure 1 Proof-of-concept demonstration of pooled screen for high-throughput functional genomics in *E. coli* based on CRISPRi technology. An sgRNA library was synthesized on a DNA microarray and consisted of four sub-libraries (Library I, II, III and NC). Library I

897 consisted of auxotrophic genes (MOPS medium) transcribed as monocistronic mRNA; Library
898 II consisted of polycistronic-mRNA-transcribing operons with one auxotrophic gene (MOPS
899 medium) in each operon; Library III consisted of genes whose knockout results in growth
900 perturbation in acidic medium; Library NC contains sgRNAs with no target site in the *E. coli*
901 genome and was used as the negative control in data processing. Oligonucleotides were
902 amplified and cloned into expression plasmids, which were transformed into *E. coli* expressing
903 dCas9 protein, resulting in relevant cell libraries. The cell libraries were grown in selective and
904 control conditions. Plasmids were extracted and deep sequencing libraries were constructed to
905 resolve the change in abundance of each sgRNA between conditions (sgRNA fitness). The
906 sgRNA fitness distribution (red histogram) of each gene was compared with that of control
907 sgRNAs (no target site in the *E. coli* genome; grey histogram) to evaluate the extent to which
908 this gene was associated with relevant phenotypes (growth perturbation under selective
909 conditions). This provided estimations of statistical significance (*P* value derived from
910 Mann-Whitney U test) and phenotypic effect (median sgRNA fitness) for each gene.
911
912 Figure 2 Genes that are auxotrophic in MOPS medium can be robustly recovered by
913 CRISPRi-based pooled screen. (a) Box plot of sgRNA *rho* score distributions for genes in
914 Libraries I and II. Outliers outside the 1.5-fold interquartile range from the box boundary were
915 plotted individually (dots). For clarity, only sgRNAs with *rho* scores between –0.4 and 0.4

were plotted. (b) $-\text{Log}_{10}$(*P* value), where *P* values were derived from Mann-Whitney U tests of sgRNA *rho* scores belonging to the indicated gene against the control sgRNA set. Filled bars represent genes with auxotrophic phenotypes reported in the Keio collection characterization; unfilled bars represent genes in Library II for which knockout does not impair growth in MOPS. Asterisks indicate $P < 0.01$ ($-\text{Log}_{10}P > 2$); red asterisks indicate false positives with significant *P* values, but with unexpected sgRNA median *rho* scores (>0) or belonging to the non-auxotrophic group (see Methods).

Figure 3 CRISPRi-based pooled screen yielded robust identification of phenotype-associated genes in polycistronic-mRNA-transcribing operons. Heat maps of $-\text{Log}_{10}$(*P* value) and median sgRNA *rho* score are shown for all nine polycistronic operons in Library II. Auxotrophic genes in MOPS medium are underlined; significant hits ($P < 0.01$, $-\text{Log}_{10}P > 2$) are indicated with asterisks. Black asterisks are true positive; red asterisks are false positives with significant *P* values but have unexpected sgRNA median *rho* scores (>0) or belong to the non-auxotrophic group (see Methods).

Figure 4 Genes with moderate growth impacts in acidic medium can be reliably recovered by CRISPRi-based pooled screen. Top: Median sgRNA *rho* score for each gene; bottom: $-\text{Log}_{10}$(*P* value) derived from Mann-Whitney U tests of sgRNA *rho* scores belonging to one

935 gene against the control sgRNA set. Asterisks indicate genes with FDR < 0.05. Hit-gene
936 calling is based on all sgRNAs belonging to one gene against the negative control sgRNA set.
937
938 Figure 5 sgRNAs residing within the first 5% of ORF region were significantly more active
939 than all other sgRNAs. Functional sgRNAs of true positive hit-genes (468 sgRNAs) in Library
940 I were grouped according to their relative position within the ORF. The absolute value of the Z
941 scores for each sgRNA were extracted and the distribution of each group against all sgRNAs in
942 this dataset was tested by the Mann-Whitney U test. Triple asterisks indicate $P < 0.01$. Results
943 are presented as box plots of absolute value of Z score for each group. Outliers outside the
944 1.5-fold interquartile range from the box boundary were plotted individually (dots). For clarity,
945 only sgRNAs with absolute Z scores between 0 and 4 were plotted.
946
947 Figure 6 Performance of pooled screen to identify phenotype-determined genes with reduced
948 sgRNA sets. Results are shown for subsampling of 3, 5, 10, 15 and 30 sgRNAs for each gene
949 (16 true positive genes in Library I). Genes without enough sgRNAs were excluded from
950 analysis when subsampling a larger number of sgRNAs than available. Hence, 16, 16, 16, 13,
951 13 and 12 genes were used when subsampling 3, 5, 10, 15, 20 and 30 sgRNAs per gene. Five
952 algorithms were applied to determine the priority of sgRNA selection during subsampling (see
953 Methods), namely Position, Cas9cal, Random, RS2, and SSC. Results are presented as box

954 plots of $-\text{Log}_{10}(P \text{ value})$ by Mann-Whitney U test. Dashed line refers to the significance

955 threshold, $\text{Log}_{10}(P \text{ value}) = 0.01$, which is used in most cases of this work.

956

957

References

1. van Opijnen T, Bodi KL, Camilli A. 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6:767–772.

2. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, Blow MJ, Bristow J, Butland G, Arkin AP, Deutschbauer A. 2015. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio 6:e00306-15.

3. Warner JR, Reeder PJ, Karimpour-Fard A, Woodruff LB a, Gill RT. 2010. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. Nat Biotechnol 28:856–62.

4. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian K-D, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang C, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. 2002.

977 Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–391.

978 5. Turner KH, Wessel AK, Palmer GC, Murray JL, Whiteley M. 2015. Essential genome

979 of Pseudomonas aeruginosa in cystic fibrosis sputum. Proc Natl Acad Sci

980 112:4110–4115.

981 6. Freed EF, Winkler JD, Weiss SJ, Garst AD, Mutalik VK, Arkin AP, Knight R, Gill RT.

982 2015. Genome-Wide Tuning of Protein Expression Levels to Rapidly Engineer

983 Microbial Traits. ACS Synth Biol 4:1244–1253.

984 7. Jiang W, Bikard D, Cox D, Zhang F, Marraffini L a. 2013. RNA-guided editing of

985 bacterial genomes using CRISPR-Cas systems. Nat Biotechnol 31:233–9.

986 8. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W,

987 Marraffini L a, Zhang F. 2013. Multiplex genome engineering using CRISPR/Cas

988 systems. Science 339:819–23.

989 9. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. 2013.

990 One-step generation of mice carrying mutations in multiple genes by

991 CRISPR/Cas-mediated genome engineering. Cell 153:910–8.

992 10. Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh J-RJ,

993 Joung JK. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat

994 Biotechnol 31:227–229.

995 11. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. 2013. Genome

engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res 41:4336–4343.

12. Bassett AR, Tibbit C, Ponting CP, Liu J-L. 2014. Mutagenesis and homologous recombination in Drosophila cell lines using CRISPR/Cas9. Biol Open 3:42–49.

13. Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, Mupo A, Grinkevich V, Li M, Mazan M, Gozdecka M, Ohnishi S, Cooper J, Patel M, McKerrell T, Chen B, Domingues AF, Gallipoli P, Teichmann S, Ponstingl H, McDermott U, Saez-Rodriguez J, Huntly BJP, Iorio F, Pina C, Vassiliou GS, Yusa K. 2016. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep 17:1193–1205.

14. Sanjana NE, Shalem O, Zhang F. 2014. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods 11:783–784.

15. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell 159:647–61.

16. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152:1173–83.

17. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, Nureki O, Zhang F. 2014. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517:583–588.

18. Otoupal PB, Erickson KE, Bordoy AE, Chatterjee A. 2016. CRISPR perturbation of gene expression alters bacterial fitness under stress and reveals underlying epistatic constraints. ACS Synth Biol acssynbio.6b00050.

19. Lv L, Ren Y-L, Chen J-C, Wu Q, Chen G-Q. 2015. Application of CRISPRi for prokaryotic metabolic engineering involving multiple genes, a case study: Controllable P(3HB-co-4HB) biosynthesis. Metab Eng 29:160–8.

20. Choudhary E, Thakur P, Pareek M, Agarwal N. 2015. Gene silencing by CRISPR interference in mycobacteria. Nat Commun 6:6267.

21. Yao L, Cengic I, Anfelt J, Hudson EP. 2016. Multiple Gene Repression in Cyanobacteria Using CRISPRi. ACS Synth Biol 5:207–212.

22. Xu T, Li Y, Shi Z, Hemme CL, Li Y, Zhu Y, Van Nostrand JD, He Z, Zhou J. 2015. Efficient Genome Editing in Clostridium cellulolyticum via CRISPR-Cas9 Nickase. Appl Environ Microbiol 81:4423–31.

23. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, Hawkins JS, Lu CHS, Koo B-M, Marta E, Shiver AL, Whitehead EH, Weissman JS, Brown ED, Qi LS, Huang

KC, Gross CA. 2016. A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. Cell 165:1493–506.

24. Struhl K. 1999. Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. Cell 98:1–4.

25. Kuzminov A. 2014. The Precarious Prokaryotic Chromosome. J Bacteriol 196:1793–1806.

26. Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. 2016. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. Elife 5:339–350.

27. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 34:184–191.

28. Chari R, Mali P, Moosburner M, Church GM. 2015. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat Methods 1–7.

29. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006.0008.

1053 30. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM,
1054 Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA.
1055 2011. Phenotypic landscape of a bacterial cell. Cell 144:143–56.
1056 31. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L,
1057 García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L,
1058 Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C,
1059 Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A,
1060 Hernández-Koutoucheva A, Del Moral-Chavez V, Rinaldi F, Collado-Vides J. 2016.
1061 RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif
1062 clustering and beyond. Nucleic Acids Res 44:D133–D143.
1063 32. Engler C, Gruetzner R, Kandzia R, Marillonnet S. 2009. Golden gate shuffling: a
1064 one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS One
1065 4:e5553.
1066 33. Kampmann M, Bassik MC, Weissman JS. 2013. Integrated platform for genome-wide
1067 screening and construction of high-density genetic interaction maps in mammalian cells.
1068 Proc Natl Acad Sci U S A 110:E2317-26.
1069 34. Mee MT, Collins JJ, Church GM, Wang HH. 2014. Syntrophic exchange in synthetic
1070 microbial communities. Proc Natl Acad Sci U S A 111:E2149-56.
1071 35. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J,

San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL. 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio 5:e01442-14.

36. Basturea GN, Rudd KE, Deutscher MP. 2006. Identification and characterization of RsmE, the founding member of a new RNA base methyltransferase family. RNA 12:426–34.

37. Chuai G-H, Wang Q-L, Liu Q. 2017. In Silico Meets In Vivo: Towards Computational CRISPR-Based sgRNA Design. Trends Biotechnol 35:12–21.

38. Farasat I, Salis HM. 2016. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. PLoS Comput Biol 12:e1004724.

39. Xu H, Xiao T, Chen C-H, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS, Brown M, Liu XS. 2015. Sequence determinants of improved CRISPR sgRNA design. Genome Res 25:1147–57.

40. Binder S, Siedler S, Marienhagen J, Bott M, Eggeling L. 2013. Recombineering in Corynebacterium glutamicum combined with optical nanosensors: a general strategy for fast producer strain generation. Nucleic Acids Res 41:6360–9.

41. Zhang Y, Buchholz F, Muyrers JPP, Stewart AF. 1998. A new logic for DNA engineering using recombination in Escherichia coli. Nature 20:123–128.

42. Walsh RM, Hochedlinger K, Braff J, Moosburner M, Yaung S, Church G, Norville J,

Church G, Jiang W, Marraffini L, Zhang F. 2013. A variant CRISPR-Cas9 system adds versatility to genome engineering. Proc Natl Acad Sci 110:15514–15515.

43. Sun Z, Deng A, Hu T, Wu J, Sun Q, Bai H, Zhang G, Wen T. 2015. A high-efficiency recombineering system with PCR-based ssDNA in Bacillus subtilis mediated by the native phage recombinase GP35. Appl Microbiol Biotechnol 99:5151–5162.

44. Gallagher LA, Ramage E, Patrapuvich R, Weiss E, Brittnacher M, Manoil C. 2013. Sequence-defined transposon mutant library of Burkholderia thailandensis. MBio 4:e00604-13.

45. Baugh L, Gallagher LA, Patrapuvich R, Clifton MC, Gardberg AS, Edwards TE, Armour B, Begley DW, Dieterich SH, Dranow DM, Abendroth J, Fairman JW, Fox D, Staker BL, Phan I, Gillespie A, Choi R, Nakazawa-Hewitt S, Nguyen MT, Napuli A, Barrett L, Buchko GW, Stacy R, Myler PJ, Stewart LJ, Manoil C, Van Voorhis WC. 2013. Combining Functional and Structural Genomics to Sample the Essential Burkholderia Structome. PLoS One 8:e53851.

46. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, Archambaud C, Cossart P, Sorek R. 2012. Comparative transcriptomics of pathogenic and non-pathogenic Listeria species. Mol Syst Biol 8:583.

47. Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, Sorek R. 2016. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science

(80-) 352:aad9822-aad9822.

48. Zhu S, Li W, Liu J, Chen C-H, Liao Q, Xu P, Xu H, Xiao T, Cao Z, Peng J, Yuan P, Brown M, Shirley Liu X, Wei W. 2016. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. Nat Biotechnol 1–10.

49. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, Mandegar MA, Olvera MP, Gilbert LA, Conklin BR, Chang HY, Weissman JS, Lim DA. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science (80-) 355.

50. Garst AD, Bassalo MC, Pines G, Lynch SA, Halweg-Edwards AL, Liu R, Liang L, Wang Z, Zeitoun R, Alexander WG, Gill RT. 2016. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat Biotechnol 35:48–55.

51. Pozsgai ER, Blair KM, Kearns DB. 2012. Modified mariner transposons for random inducible-expression insertions and transcriptional reporter fusion insertions in Bacillus subtilis. Appl Environ Microbiol 78:778–85.

52. Oakes BL, Nadler DC, Flamholz A, Fellmann C, Staahl BT, Doudna JA, Savage DF. 2016. Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. Nat Biotechnol.

53. Heigwer F, Kerr G, Boutros M. 2014. E-CRISP: fast CRISPR target site identification. Nat Methods 11:122–123.

54. Liu H, Wei Z, Dominguez A, Li Y, Wang X, Qi LS. 2015. CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. Bioinformatics 31:3676–3678.

55. Connell MRO, Oakes BL, Sternberg SH, East-seletsky A, Kaplan M, Doudna JA. 2014. Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature 516:263–266.

56. Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, Doudna JA, Yeo GW. 2016. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell 165:488–496.

57. Liu Y, Chen Z, He A, Zhan Y, Li J, Liu L, Wu H, Zhuang C, Lin J, Zhang Q, Huang W. 2016. Targeting cellular mRNAs translation by CRISPR-Cas9. Sci Rep 6:29652.

58. Goh S, Hohmeier A, Stone TC, Offord V, Sarabia F, Garcia-Ruiz C, Good L. 2015. Silencing of Essential Genes within a Highly Coordinated Operon in Escherichia coli. Appl Environ Microbiol 81:5650–5659.

59. Smith JD, Suresh S, Schlecht U, Wu M, Wagih O, Peltz G, Davis RW, Steinmetz LM, Parts L, St Onge RP. 2016. Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. Genome Biol 17:45.

60. Cui L, Bikard D. 2016. Consequences of Cas9 cleavage in the chromosome of

Escherichia coli. Nucleic Acids Res 44:4243–4251.

61. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. 2013. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res 41:7429–37.

62. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6:343–5.

63. Neidhardt FC, Bloch PL, Smith DF. 1974. Culture medium for enterobacteria. J Bacteriol 119:736–47.

64. Murphy KC, Campellone KG, Poteete AR. 2000. PCR-mediated gene replacement in Escherichia coli. Gene 246:321–330.

65. Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. 2015. Multigene editing in the Escherichia coli genome using the CRISPR-Cas9 system. Appl Environ Microbiol 81:AEM.04023-14.

66. Zhao J, Li Q, Sun T, Zhu X, Xu H, Tang J, Zhang X, Ma Y. 2013. Engineering central metabolic modules of Escherichia coli for improving β-carotene production. Metab Eng 17:42–50.

67. Amann E, Ochs B, Abel KJ. 1988. Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. Gene 69:301–315.

1167 68. Lutz R, Bujard H. 1997. Independent and tight regulation of transcriptional units in
1168 escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.
1169 Nucleic Acids Res 25:1203–1210.
1170 69. Salis HM, Mirsky EA, Voigt CA. 2009. Automated design of synthetic ribosome
1171 binding sites to control protein expression. Nat Biotechnol 27:946–50.
1172 70. Knaus R, Bujard H. 1988. PL of coliphage lambda: an alternative solution for an
1173 efficient promoter. EMBO J 7:2919–2923.
1174 71. Ronda C, Pedersen LE, Sommer MOA, Nielsen AT. 2016. CRMAGE: CRISPR
1175 Optimized MAGE Recombineering. Sci Rep 6:19452.
1176 72. Jiang H, Wong WH. 2008. SeqMap: Mapping massive amount of oligonucleotides to
1177 the genome. Bioinformatics 24:2395–2396.
1178 73. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ,
1179 Wu X, Shalem O, Cradick TJ, Marraffini LA, Bao G, Zhang F. 2013. DNA targeting
1180 specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 31:827–32.
1181 74. Sander JD, Joung JK. 2014. CRISPR-Cas systems for editing, regulating and targeting
1182 genomes. Nat Biotechnol 32:347–355.
1183 75. Li W, Köster J, Xu H, Chen C-H, Xiao T, Liu JS, Brown M, Liu XS. 2015. Quality
1184 control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR.
1185 Genome Biol 16:281.

Author Contributions

T.W., C.Z. and X.X. proposed the general design of this work. T.W. and J.G. performed the experiments. T.W. and C.G. performed the data processing. Y.W. helped with plasmid construction. B.L. prepared the plasmid library. T.W., J.G., Z.X., C.Z. and X.X. analysed the results. T.W., J.G. and C.Z. wrote the manuscript based on discussion among all authors.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC21627812) and Tsinghua University Initiative Scientific Research Program (20161080108).

Availability of data and materials

The authors declare that all data supporting the findings of this study are available within the article and other files such as homemade scripts are available from the corresponding author upon request.

Figure 1

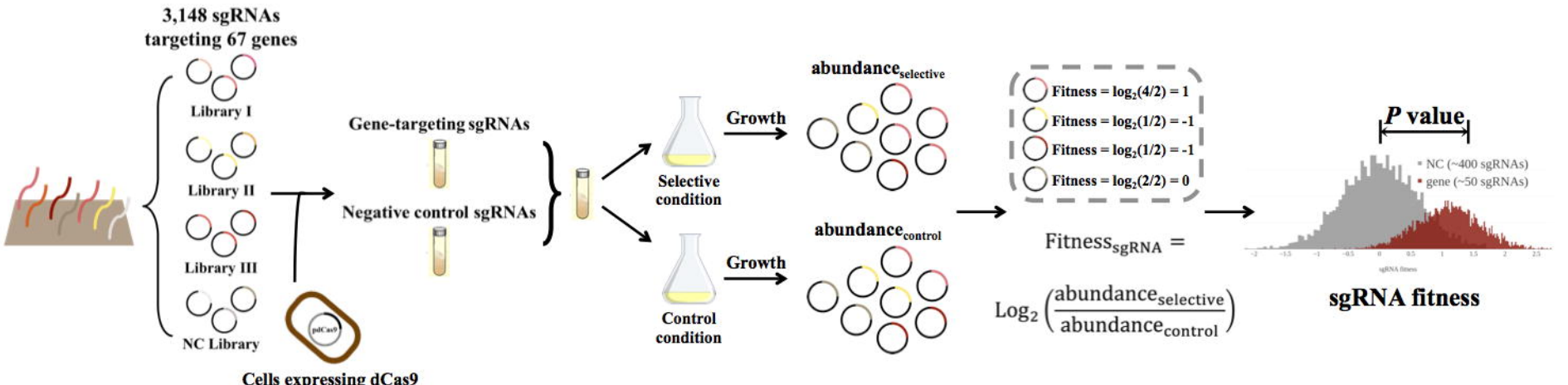
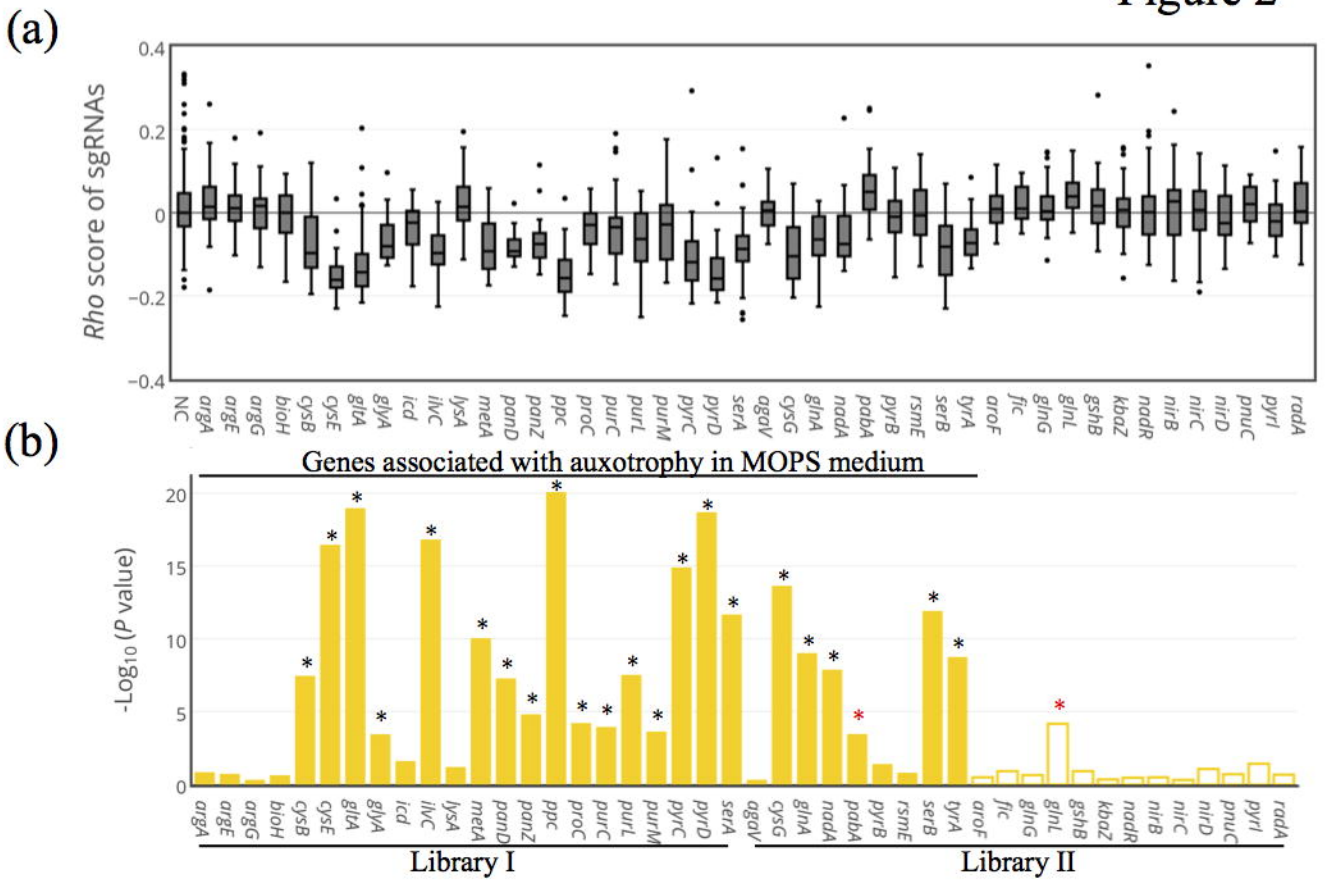

Library design
Library construction
Selection
sgRNA fitness calculation
Gene fitness calculation
3,148 sgRNAs
targeting 67 genes
Library I
Library II
Library III
NC Library
Cells expressing dCas9
Gene-targeting sgRNAs
Negative control sgRNAs
Selective condition
Control condition
Growth
Growth
abundance selective
abundance control
Fitness = log2(4/2) = 1
Fitness = log2(1/2) = -1
Fitness = log2(1/2) = -1
Fitness = log2(2/2) = 0
Fitness sgRNA =
Log2 (abundance selective / abundance control)
P value
NC (~400 sgRNAs)
gene (~50 sgRNAs)
sgRNA fitness

Figure 2


(a)
Rho score of sgRNAs
0.4
0.2
0
−0.2
−0.4
NC argA argE argG bioH cysB cysE gltA glyA icd ilvC lysA metA panD panZ ppc proC purC purL purM pyrC pyrD serA agaV cysG glnA nadA pabA pyrB rsmE serB tyrA aroF fic glnG glnL gshB kboZ nadR nirB nirC nirD pnuC pyrI radA
(b)
Genes associated with auxotrophy in MOPS medium
-Log10 (P value)
20
15
10
5
0
argA argE argG bioH cysB cysE gltA glyA icd ilvC lysA metA panD panZ ppc proC purC purL purM pyrC pyrD serA
Library I
agaV cysG glnA nadA pabA pyrB rsmE serB tyrA aroF fic glnG glnL gshB kboZ nadR nirB nirC nirD pnuC pyrI radA
Library II

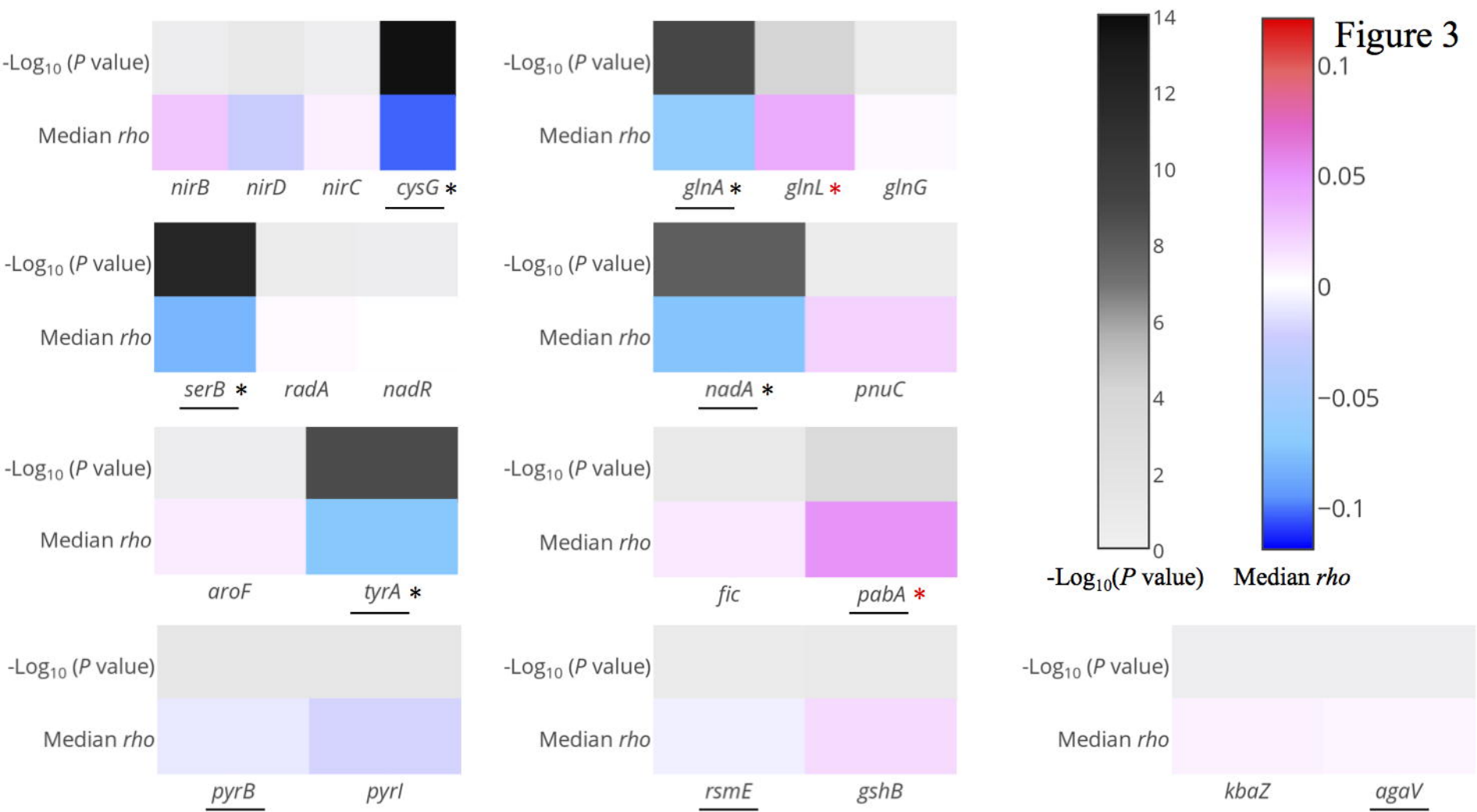
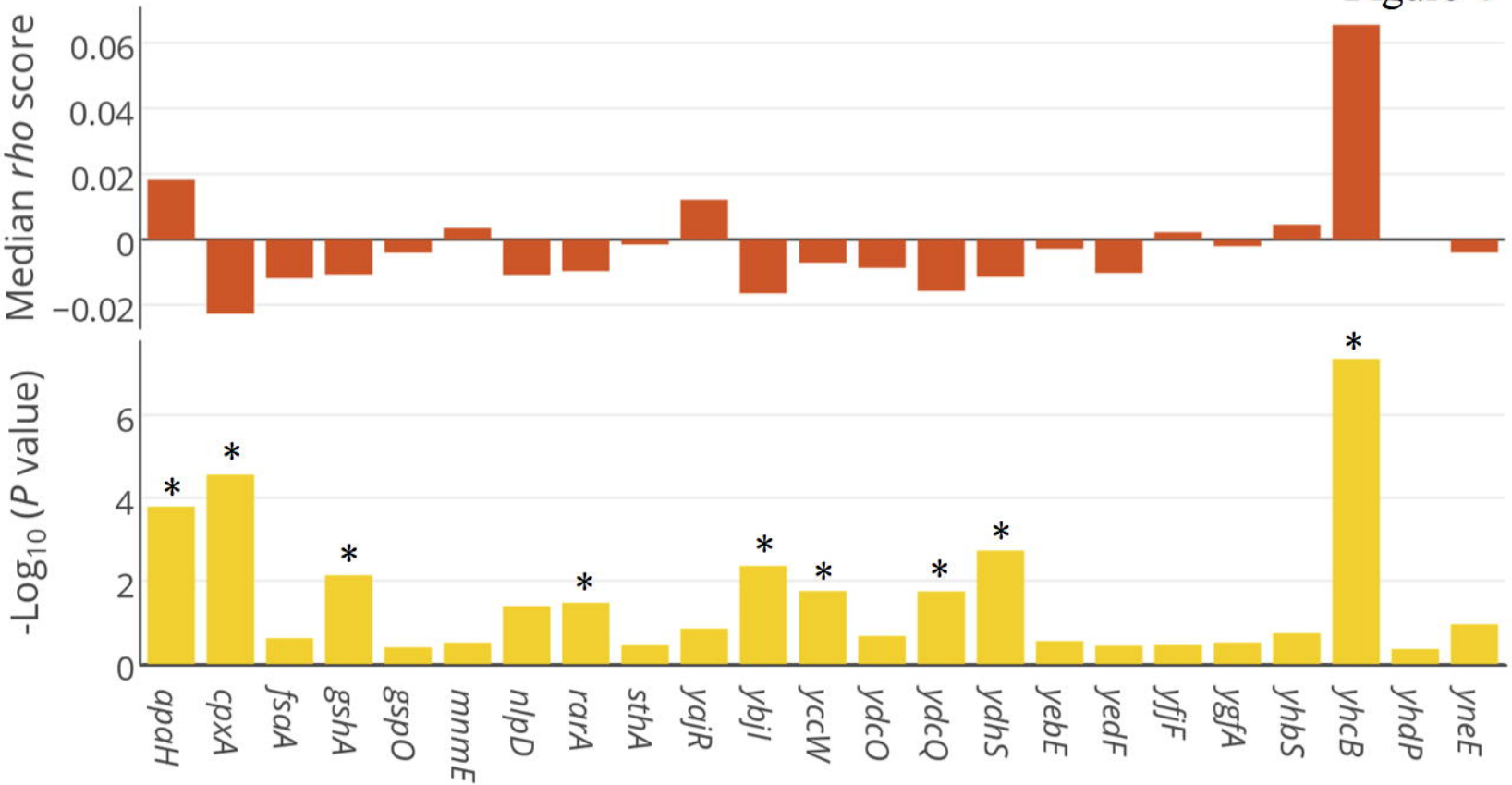

-Log10 (P value)
Median rho
nirB
nirD
nirC
cysG *
glnA *
glnL *
glnG
serB *
radA
nadR
nadA *
pnuC
aroF
tyrA *
fic
pabA *
pyrB
pyrI
rsmE
gshB
kbaZ
agaV
14
12
10
8
6
4
2
0
0.1
0.05
0
−0.05
−0.1
-Log10(P value)
Median rho

Figure 3

Figure 4


Median rho score
0.06
0.04
0.02
0
−0.02
-Log10 (P value)
6
4
2
0
*
apaH
cpxA
fsaA
gshA
gspO
mnmE
nlpD
rarA
sthA
yajR
ybjI
yccW
ydcO
ydcQ
ydhS
yebE
yedF
yfjF
ygfA
yhbS
yhcB
yhdP
yneE

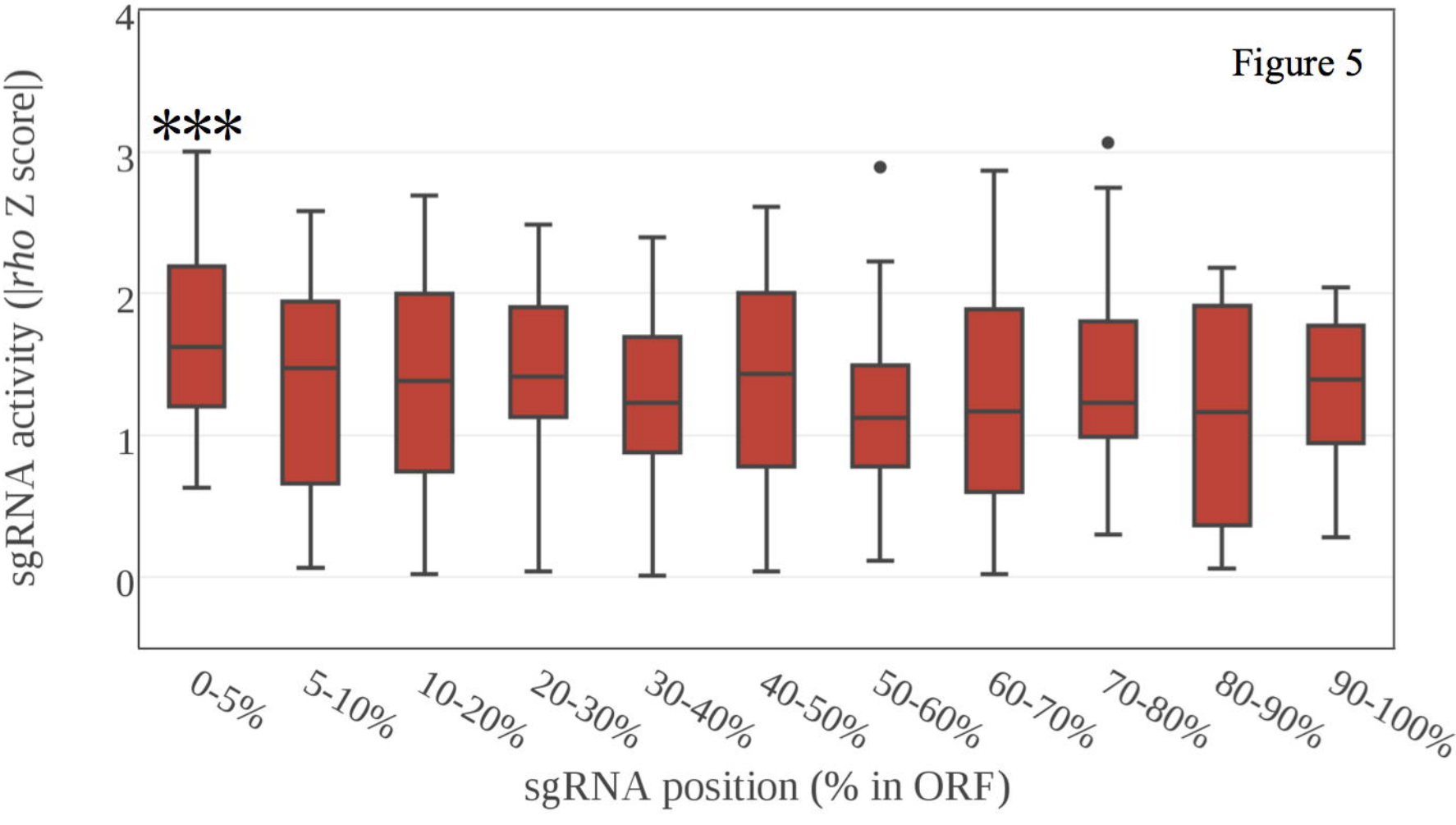


Figure 5
***
sgRNA activity (|rho Z score|)
4
3
2
1
0
0-5%
5-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
90-100%
sgRNA position (% in ORF)

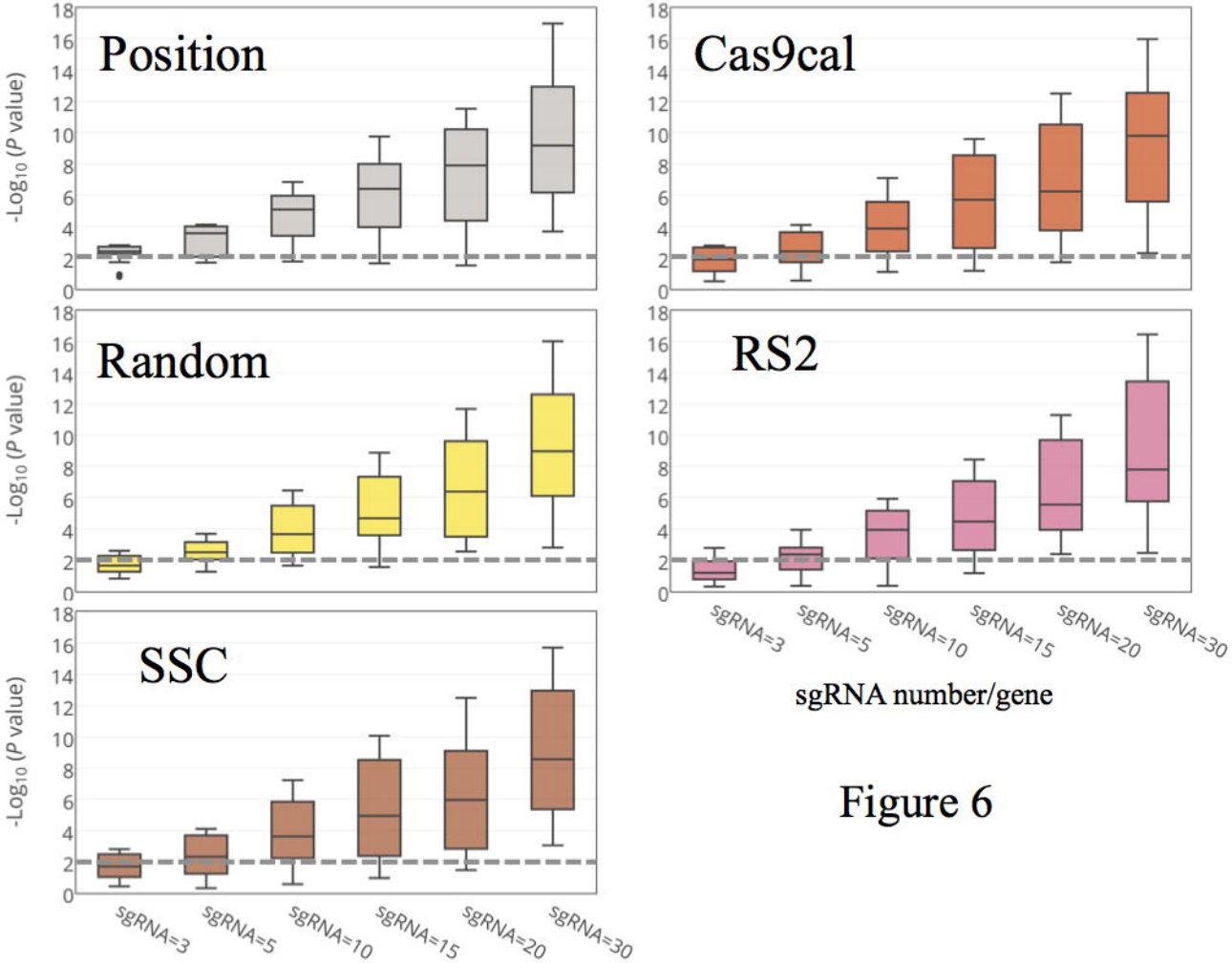

Position
Cas9cal
Random
RS2
SSC
-Log10 (P value)
sgRNA=3
sgRNA=5
sgRNA=10
sgRNA=15
sgRNA=20
sgRNA=30
sgRNA number/gene

Figure 6