

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34

## **Dub-seq: dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits**

Vivek K. Mutalik<sup>1,\*</sup>, Pavel S. Novichkov<sup>1</sup>, Morgan N. Price<sup>1</sup>, Trenton K. Owens<sup>1</sup>, Mark Callaghan<sup>1</sup>, Sean Carim<sup>2</sup>, Adam M. Deutschbauer<sup>1,2</sup>, Adam P. Arkin<sup>1,3,\*</sup>

<sup>1</sup>Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory

<sup>2</sup>Department of Plant and Microbial Biology, University of California, Berkeley

<sup>3</sup>Department of Bioengineering, University of California, Berkeley

\*To whom correspondence should be addressed:

[vkmutalik@lbl.gov](mailto:vkmutalik@lbl.gov); [aparkin@lbl.gov](mailto:aparkin@lbl.gov)

## 35 **Abstract**

36 A major challenge in genomics is the knowledge gap between sequence and its  
37 encoded function. Gain-of-function methods based on gene overexpression are  
38 attractive avenues for phenotype-based functional screens, but are not easily applied in  
39 high-throughput across many experimental conditions. Here, we present **Dual Barcoded**  
40 **Shotgun Expression Library Sequencing** (Dub-seq), a method that greatly increases the  
41 throughput of genome-wide overexpression assays. In Dub-seq, a shotgun expression  
42 library is cloned between dual random DNA barcodes and the precise breakpoints of  
43 DNA fragments are associated to the barcode sequences prior to performing assays. To  
44 assess the fitness of individual strains carrying these plasmids, we use DNA barcode  
45 sequencing (BarSeq), which is amenable to large-scale sample multiplexing. As a  
46 demonstration of this approach, we constructed a Dub-seq library with total *Escherichia*  
47 *coli* genomic DNA, performed 155 genome-wide fitness assays in 52 experimental  
48 conditions, and identified 813 genes with high-confidence overexpression phenotypes  
49 across 4,151 genes assayed. We show that Dub-seq data is reproducible, accurately  
50 recapitulates known biology, and identifies hundreds of novel gain-of-function  
51 phenotypes for *E. coli* genes, a subset of which we verified with assays of individual  
52 strains. Dub-seq provides complementary information to loss-of-function approaches  
53 such as transposon site sequencing or CRISPRi and will facilitate rapid and systematic  
54 functional characterization of microbial genomes.

55

56

## 57 **Importance**

58 Measuring the phenotypic consequences of overexpressing genes is a classic genetic  
59 approach for understanding protein function; for identifying drug targets, antibiotic and  
60 metal resistance mechanisms; and for optimizing strains for metabolic engineering. In  
61 microorganisms, these gain-of-function assays are typically done using laborious  
62 protocols with individually archived strains or in low-throughput following qualitative  
63 selection for a phenotype of interest, such as antibiotic resistance. However, many  
64 microbial genes are poorly characterized and the importance of a given gene may only  
65 be apparent under certain conditions. Therefore, more scalable approaches for gain-of-  
66 function assays are needed. Here, we present **Dual Barcoded Shotgun Expression**  
67 **Library Sequencing** (Dub-seq), a strategy that couples systematic gene overexpression  
68 with DNA barcode sequencing for large-scale interrogation of gene fitness under many  
69 experimental conditions at low cost. Dub-seq can be applied to many microorganisms  
70 and is a valuable new tool for large-scale gene function characterization.

71

## 72 INTRODUCTION

73  
74 Advances in DNA sequencing have had a tremendous impact on microbial genomics,  
75 as thousands of genomes have now been sequenced<sup>1</sup>. However, only a small fraction  
76 of these microorganisms have been experimentally studied and as such, our predictions  
77 of gene function, metabolic capability, and community function for these  
78 microorganisms are based largely on automated computational approaches<sup>2</sup>.  
79 Unfortunately, many of these computational predictions are incomplete or erroneous,  
80 especially in instances where the homology of a sequenced gene is too distant from any  
81 experimentally characterized relative<sup>3</sup>. To bridge this gap between sequencing and  
82 functional characterization, it is imperative that large-scale, inexpensive, and organism-  
83 agnostic tools are developed and applied<sup>4</sup>.

84  
85 A number of large-scale approaches based on loss-of-function genetics have been  
86 developed for microorganisms including gene-knockout libraries<sup>5-9</sup>, recombineering  
87 based methods<sup>10,11</sup>, transposon mutagenesis coupled to next-generation sequencing  
88 (TnSeq)<sup>12,13</sup>, and CRISPR interference (CRISPRi)<sup>14</sup>. Collectively, these strategies all  
89 rely on measuring the phenotypic consequences of removing a gene from a  
90 microorganism and inferring protein function based on these phenotypes. An  
91 adaptation of TnSeq that incorporates and uses random DNA barcodes (RB-TnSeq) to  
92 measure strain abundance in a competitive growth assay<sup>13</sup> has recently been applied  
93 on a larger scale to identify mutant phenotypes for thousands of genes across 32  
94 bacteria<sup>15</sup>. Despite their utility, these loss-of-function approaches suffer some  
95 limitations: only CRISPRi is effective for interrogating essential genes under multiple  
96 conditions, it is challenging to identify phenotypes for genes with redundant functions  
97 using single mutants, and these approaches require some degree of genetic tractability  
98 in the target microorganism.

99  
100 A complimentary approach for studying gene and organism function is to generate gain-  
101 of-function overexpression libraries and analyze the phenotypic consequences of  
102 increased gene dosage. Indeed, the impact of enhanced gene dosage on adaptation  
103 and evolution are well documented across all three kingdoms of life and have been  
104 shown to be an important contributor to numerous diseases and drug-resistance  
105 phenotypes<sup>16-18</sup>. Overexpression as a genetic tool has a rich history of connecting  
106 genes to cellular functions and has been exploited as a versatile screening technique to  
107 identify drug targets<sup>16,19,20</sup>, antibiotic and metal resistance genes<sup>17,21,22</sup>, virus-resistance  
108 genes<sup>23</sup>, genetic suppressors<sup>24,25</sup>, as well as for a number of chemical genomics<sup>8,9</sup> and  
109 biotechnology applications<sup>26-28</sup>. While a number of technologies have been developed  
110 for overexpression screens including defined open reading frame (ORF) libraries<sup>6,20,29</sup>  
111 and activation modes of recombineering<sup>30,31</sup>, transposon insertions<sup>32</sup> or CRISPR

112 systems<sup>33</sup>, these strategies are limited, either due to the need for expensive and  
113 laborious generation of archived strains or the need for organism-specific genetic tools.

114

115 A simpler alternative for overexpression screens is a shotgun library-based approach in  
116 which random DNA is introduced into a host organism for phenotyping and functional  
117 assessment. This approach has been widely used for studying increased-copy number  
118 effects on a desired phenotype<sup>26,27</sup> and for activity-based screening of metagenomic  
119 samples<sup>34,35</sup>. Nevertheless, most shotgun expression libraries have only been assayed  
120 in a small number of conditions looking for a specific gene-function, and are often  
121 performed as qualitative selections on a plate<sup>34-36</sup>. Furthermore, current shotgun-based  
122 approaches typically require tedious and expensive sequencing and sample preparation  
123 protocols for identifying the selected gene(s)<sup>26,27,37,38</sup>. With arrival of next-generation  
124 sequencing technologies, all positive candidates can be pooled, and cloned regions can  
125 be amplified and sequenced in parallel<sup>39,40</sup>. Unfortunately, sequencing the cloned  
126 regions (to identify the genes conferring the phenotype) is labor intensive and may  
127 become cost-prohibitive if the overexpression library is being assayed in many  
128 conditions. As such, there is a need for high-throughput gain-of-function technology that  
129 is simple, quantitative, agnostic to source DNA, and which facilitates multiplexed  
130 quantification of fitness under hundreds of experimental conditions.

131

132 Here we present a new method termed Dub-seq, or dual barcoded shotgun expression  
133 library sequencing, for performing high-throughput and quantitative gain-of-function  
134 screens. Dub-seq requires an initial characterization of the overexpression library by  
135 linking the genomic breakpoints of each clone to a pair of random DNA barcodes.  
136 Subsequent screens are performed using a competitive fitness assay with a simple  
137 DNA barcode sequencing and quantification assay (BarSeq<sup>41</sup>). As a demonstration of  
138 this approach, we generated an *E. coli* Dub-seq library and assayed the phenotypic  
139 consequences of overexpressing nearly all genes on *E. coli* fitness under dozens of  
140 experimental conditions. We show that Dub-seq yields gene fitness data that is  
141 consistent with known biology and also provides novel gene-function insights. We  
142 validate some of these new findings by overexpressing individual genes and quantifying  
143 these strains' fitness. Given that only DNA and a suitable host organism for assaying  
144 fitness are necessary, Dub-seq can be readily extended to diverse functional genomics  
145 and biotechnology applications.

146

147

148

149

150

151

## 152 RESULTS

153

### 154 Overview of Dub-seq

155 The Dub-seq approach is summarized in Figure 1 and can be separated into four different  
156 steps. First, a plasmid library is generated with pairs of random 20 nucleotide DNA  
157 sequences, termed the UP and DOWN barcodes. To link the identities of the two-  
158 barcode sequences on each plasmid, Barcode-Pair sequencing (BPseq) is performed  
159 (Fig. 1a, Methods). Second, sheared genomic DNA from an organism under  
160 investigation is cloned between the previously associated UP and DOWN barcodes  
161 (Fig. 1b). Third, the genomic fragment endpoints are mapped and associated with the  
162 two-barcode sequences using a TnSeq-like protocol<sup>13</sup>. We term this step Barcode-  
163 Association-with Genome fragment by sequencing or BAGseq and the resulting plasmid  
164 library as the “Dub-seq” library (Fig. 1c). The BAGseq step requires two sample  
165 preparations to separately map genomic fragment junctions to the UP and DOWN  
166 barcodes. The BAGseq characterization generates a table of barcode sequences and  
167 the cloned chromosomal breakpoints at single-nucleotide resolution. Because the two  
168 random DNA barcodes have been previously associated, we can infer the exact  
169 sequence of each plasmid in the Dub-seq library if the sequence of the source DNA is  
170 known. Lastly, we introduce the Dub-seq plasmid library into a host bacterium and  
171 monitor the fitness of strains carrying these plasmids in a competitive fitness assay  
172 under a particular condition by PCR amplifying and quantifying the abundance of the  
173 DNA barcode sequences (BarSeq<sup>41</sup>, Fig. 1d). In these pooled fitness experiments, the  
174 barcode abundance changes depending upon the fitness phenotype imparted by the  
175 barcode-associated-genome fragments. A data analysis pipeline yields fitness scores  
176 for individual strains (or “fragments”) and for each gene. These gene scores provide an  
177 assessment of the phenotypic consequence of overexpressing nearly all of the genes  
178 represented in the cloned DNA fragments. The advantage of Dub-seq is that it  
179 decouples the characterization of a shotgun overexpression library (which is more  
180 laborious) from the cheaper and simpler fitness determination step using BarSeq. As  
181 such, a Dub-seq library can be readily assayed in hundreds of different experimental  
182 conditions. Dub-seq can be viewed as an overexpression-based, gain-of-function  
183 version of our previously described method for random barcode transposon-site  
184 sequencing (RB-TnSeq)<sup>13</sup>.

185

### 186 Generation of *E. coli* Dub-seq library

187 To generate a Dub-seq library, we used a broad host range vector with a pBBR1  
188 replication origin. We used standard molecular biology techniques to insert two random  
189 20 nucleotide barcode sequences on the plasmid, the UP and DOWN barcodes, that  
190 juxtapose a unique PmlI restriction enzyme site on the plasmid. Both the UP barcodes  
191 and DOWN barcodes contain common PCR priming sites for rapid amplification of all

192 barcodes from a pooled sample. We generated a dual barcoded vector library with  
193 ~250,000 clones in *E. coli* and characterized this library by associating the barcode  
194 pairs using BPseq. The vector library of ~250,000 clones was sufficient to map unique  
195 barcode-pairs with confidence and also to yield a Dub-seq library in which each  
196 fragment will have a unique barcode (see below).

197  
198 To generate the *E. coli* Dub-seq library, we extracted *E. coli* (BW25113) genomic DNA,  
199 sheared to 3 kb fragment size, and cloned the fragments into the dual barcoded  
200 backbone vector digested with PmlI. The *E. coli* Dub-seq library encompasses ~40,000  
201 vectors, corresponding to about 8X coverage of the *E. coli* genome. In this study, we  
202 used the endogenous *E. coli* transcription and translation apparatus to drive the  
203 expression of the encoded gene(s) within each genomic fragment, although future  
204 studies could use inducible systems (for example, when the source of the cloned Dub-  
205 seq DNA differs from the host bacterium for assaying fitness<sup>42</sup>).

206  
207 We next characterized the *E. coli* Dub-seq library using BAGseq, which identifies the  
208 cloned genome fragment and its pairings with the neighboring dual barcodes. As there  
209 are two barcodes for each Dub-seq library, we performed two separate BAGseq sample  
210 preparation steps, one for the UP barcodes and one for the DOWN barcodes. Briefly,  
211 BAGseq involves shearing of the Dub-seq plasmid library, end repair, Illumina adaptor  
212 ligation, PCR amplification of the junction between the barcode and genomic insert  
213 using primers that are complementary to one of the barcode-specific primer binding  
214 sites, and deep sequencing of these samples (modified from reference 11). After  
215 filtering out barcodes that mapped to more than one genomic fragment, we identified  
216 30,558 unique barcode pairs that we could confidently associate with a genomic  
217 fragment.

218  
219 In the *E. coli* Dub-seq library, the fragments are evenly distributed across the  
220 chromosome (**Fig. 2a**), the average fragment size is 2.6 kB (**Fig. 2b**), and the majority  
221 of fragments covered 2-3 genes in their entirety (**Fig. 2c**). 80% of genes in the *E. coli*  
222 genome are covered (from start to stop codon) by at least 5 independent genomic  
223 fragments in the Dub-seq library (**Fig. 2d**) and 97% of all genes are covered by at least  
224 one fragment. Just 135 genes are not covered in their entirety by any Dub-seq fragment  
225 (**Supplementary Table 1**). Many of these unmapped or uncovered genes encode  
226 membrane and ribosomal proteins and probably reflect the lethality of overexpressing  
227 these genes<sup>43</sup>. Other genes could not be confidently mapped because they are  
228 associated with repetitive regions. For example, we could not confidently map  
229 fragments covering ETT2 type III secretion system pathogenicity island and its regulator  
230 gene *ygeH* which has tetratricopeptide repeat motifs, while the neighboring protein-  
231 coding genes are well mapped (**Fig. 2a**). Similarly, we could not map genes within

232 ribosomal RNA operons (example, *rrlD*, **Fig. 2a**), as *E. coli* encodes multiple nearly-  
233 identical copies of these loci. Some large genes with length more than 3.5 Kb, such as  
234 *rpoB*, are not entirely covered by any fragments in our library, while other large genes  
235 such as *acrB* are covered by only one fragment (**Fig. 2a**).

236  
237 Of the *E. coli* protein-coding genes that are essential for viability when deleted<sup>5</sup>, 95%  
238 are completely covered by at least one fragment in the Dub-seq library (**Supplementary**  
239 **Table 2**). This demonstrates that the Dub-seq approach can interrogate genes that are  
240 not typically assayed for conditional phenotypes in loss-of-function approaches. There  
241 are only 17 protein-coding genes that are both essential for viability when deleted and  
242 absent from our Dub-seq library (**Supplementary Table 2**).

### 243 244 **Strain and gene fitness profiling using BarSeq**

245 The key advantage of Dub-seq is the ease of assessing the relative fitness contributions  
246 of all genes contained in the cloned genomic fragments using pooled, competitive  
247 growth assays. Depending on the assay condition and the gene(s) encoded by a  
248 genomic fragment, the relative abundance of a strain carrying that fragment can change  
249 due to its fitness advantage or disadvantage relative to strains carrying other fragments.  
250 Because the DNA barcodes have been previously associated to each genomic  
251 fragment, we can simply compare the relative abundance of each barcode before and  
252 after selective growth using DNA barcode sequencing or BarSeq<sup>41</sup>.

253  
254 As a demonstration of Dub-seq fitness assays and to illustrate our approach for  
255 calculating strain (fragment) and gene fitness scores, we recovered an aliquot of the *E.*  
256 *coli* Dub-seq library in LB to mid-log phase, collected a cell pellet for the “start” (or time-  
257 zero sample), and used the remaining cells to inoculate an LB culture supplemented  
258 with 1.2 mM nickel. After growth in the presence of nickel, we collected a second cell  
259 pellet for the “condition” sample. We extracted plasmid DNA from the start and condition  
260 samples, PCR amplified the UP and DOWN DNA barcodes from each, and sequenced  
261 the DNA barcodes with Illumina. We calculate the fragment fitness score for each strain  
262 by taking the normalized log<sub>2</sub> ratio of the number of reads for each barcode in condition  
263 sample versus the start sample (**Fig. 1**). Positive scores indicate that the gene(s)  
264 contained on that fragment lead to an increase in relative fitness, while negative values  
265 mean the gene(s) on the fragment reduced relative fitness. Scores near zero indicate no  
266 fitness reduction or benefit for the gene(s) under the assayed condition. As in previous  
267 work<sup>44</sup>, we find that fitness scores calculated with either UP barcodes or DOWN  
268 barcodes yield very similar results ( $r = 0.94$ , **Supplementary Fig. 1ab**). Therefore, we  
269 only sequenced the UP barcodes for all additional experiments in this study.

270

271 Given that multiple, causative and non-causative genes can be contained on a single  
272 fragment, to assign a fitness score to a particular gene it is necessary to examine the  
273 score of all fragments containing the gene. Here, we considered two different ways to  
274 estimate fitness score of a gene. The first approach was to simply take the average of  
275 all fitness scores for fragments that contained the gene in its entirety (the “mean” score).  
276 The second approach was to use a regression method for estimating gene fitness score  
277 so as to prevent genes from having artifactually high fitness scores if they were located  
278 near other causative genes. Specifically, we adopted non-negative least squares  
279 regression (the “regression” score) (see Methods). To illustrate how the mean and  
280 regression scores differ in practice, consider the gene fitness scores for two adjacent  
281 genes under elevated nickel stress, *rcnA* and *rcnR* (**Fig. 3a and 3b**). RcnA is a nickel  
282 efflux protein whose overexpression is known to lead to increased nickel tolerance<sup>45</sup>.  
283 Conversely, *rcnR* encodes a transcriptional repressor that weakly represses its own  
284 expression and that of *rcnA*, and the overexpression of *rcnR* alone is not expected to  
285 increase nickel tolerance<sup>45</sup>. While the mean and regression approaches both result in  
286 similar (and correct) high Dub-seq scores for *rcnA* (**Fig. 3a**), only the regression  
287 approach results in the correct, neutral fitness score for the *rcnR* (**Fig. 3b**). The mean  
288 score calculation approach leads to an artifactually high fitness score for *rcnR* because  
289 many of the fragments that contain this gene also contain the neighboring *rcnA* (**Fig. 3b**,  
290 **Supplementary Figs. 2ab and 3ab**). Based on these results and other examples  
291 (**Supplementary Fig. 4**) that we examined, we concluded that the optimal strategy was  
292 to use the regression method for calculating Dub-seq gene fitness scores (Methods).

293  
294 To assess the reproducibility of Dub-seq fitness assays, we compared the results  
295 obtained from independent samples. First, the number of sequencing read counts for  
296 each UP barcodes from the Dub-seq library from different start samples were highly  
297 correlated (**Supplementary Fig. 1c**). Likewise, between two biological replicates of the  
298 nickel stress experiment, we found a strong correlation for fragment fitness ( $r = 0.80$ ;  
299 **Fig. 3c**) and for regression-based gene fitness ( $r = 0.89$ ; **Fig. 3d**).

### 301 **Fitness profiling across dozens of experimental conditions**

302 To demonstrate the scalability of Dub-seq, we performed 155 genome-wide pooled  
303 fitness experiments representing 52 different chemicals: 23 compounds as the sole  
304 source of carbon in a defined growth media and varying concentrations of 29 inhibitory  
305 compounds in rich media (**Fig. 4**). The inhibitory compounds included metals, salts, and  
306 antibiotics. For each of these assays, we compared the abundance of the UP barcodes  
307 before and after growth selection. We multiplexed 48 or 96 BarSeq PCR samples per  
308 lane of Illumina sequencing, at a sequencing cost of about \$20 per genome-wide assay.  
309 In the typical condition sample, we obtained ~4.2 million BarSeq reads, representing  
310 ~100 reads on an average for each clone in the Dub-seq plasmid library. We computed



311 gene fitness scores (using the regression approach) for 4,027 protein-coding genes and  
312 for 124 RNA genes. The gene fitness scores were reproducible, with a median pairwise  
313 correlation of 0.80 across 64 biological replicates.

314

315 We focused on the genes with positive fitness scores, as the overexpression of a gene  
316 that is important for a given process is usually expected to lead to a fitness  
317 advantage<sup>17,46</sup>, but we also examined the negative scores. To identify a subset of the  
318 effects that were likely to be reliable, we used three filters: the fitness effect was large  
319 relative to the variation between start samples ( $|\text{score}| \geq 2$ ); the fragments containing  
320 the gene showed consistent fitness across replicate experiments (using a  $t$  test); and  
321 the number of reads for those fragments was sufficient for the gene score to have little  
322 noise (see Methods). Effects that passed these filters were more likely to be consistent  
323 in replicate experiments (for example, see **Fig. 3d**). We considered an effect that  
324 passed these filters to be of high confidence if it was based on more than one fragment  
325 or if the gene had a large effect in another experiment for the compound. Overall, we  
326 identified 4,051 high-confidence effects, representing 813 of the 4,151 genes assayed  
327 (**Supplementary Table 3**). 400 different genes had a high-confidence fitness benefit  
328 when overexpressed in at least one condition, while the overexpression of 571 different  
329 genes led to a decrease in fitness in at least one condition. Nearly all experiments (153  
330 of 155) had at least one gene with a high-confidence effect. By shuffling the  
331 measurements for each fragment in each experiment, we estimated a false discovery  
332 rate of less than 2% (Methods). Among the *E. coli* genes essential for viability when  
333 deleted<sup>5</sup>, 46 have a high-confidence benefit in at least in one experiment, demonstrating  
334 that gain-of-function approaches like Dub-seq can identify conditional phenotypes for  
335 genes that are not typically interrogated by loss-of-function approaches such as Tn-seq.

336

337 Some genes had positive fitness benefits across many conditions. In particular, five  
338 genes (*recA*, *galE*, *dgt*, *rcnA*, *fabB*) had high-confidence benefits in 10 or more different  
339 conditions. The most frequent benefits were found for *recA* and *galE*, which are  
340 disrupted in the DH10B derivative host strain we used<sup>47</sup> (Methods). Even for pleiotropic  
341 genes, we find that they confer a more extreme beneficial phenotype in some  
342 conditions. For example, UDP-glucose 4-epimerase (*galE*) is highly beneficial to  
343 overexpress in the presence of 0.1 mM benzethonium chloride, with gene scores of +12  
344 or +14 in two replicate experiments. All of *galE*'s other scores were under +5. Similarly,  
345 strand exchange and recombination gene *recA* shows high fitness scores of +6 in the  
346 presence of cisplatin, lomefloxacin and sodium chloride. In addition to these examples,  
347 we found that 32 genes provide growth advantage in 5 or more antibiotics, metals or  
348 other stress conditions, as compared to 241 genes showing growth benefit in just one  
349 condition (**Supplementary Table 3**).

350

351 Some of the Dub-seq experiments identified dozens of putatively beneficial genes. For  
352 example, with potassium acetate as the carbon source, we identified 56 genes that had  
353 high-confidence benefits in both of two replicate experiments (**Supplementary Table**  
354 **3**). The two highest-scoring genes encode isozymes of aconitase (*acnA* and *acnB*),  
355 which are part of the tricarboxylic acid cycle for oxidizing acetate<sup>48</sup>. But the relationship  
356 between the other beneficial genes and acetate catabolism is not obvious. As another  
357 example, in copper (II) chloride stress at 2 mM, 120 genes had high-confidence  
358 benefits. The genes with the highest scores were *envZ*, *mltD*, *citB/dpiA*, *mepM*, *mepS*,  
359 *cutC*, and other high-scoring genes encode outer membrane porins (*ompX*, *ompC*,  
360 *ompF*) or lipoprotein *nlpE* (**Supplementary Table 3**). Overexpression of most these  
361 genes is known to activate the complex regulatory network of envelope stress response  
362 via *cpxAR* and sigma-E<sup>49,50</sup>. Specifically, it is known that the copper tolerance  
363 phenotype observed in the case of *nlpE* overexpression is due to activation of Cpx  
364 pathway<sup>51</sup>. In the case of *cutC* overexpression, sigma-E driven small RNA *micL*  
365 encoded within *cutC* is overproduced, leads to targeted downregulation of *lpp* and  
366 sufficient for copper tolerance phenotype<sup>52</sup>. Finally, dozens of genes show growth  
367 benefits in the presence of the membrane-disrupting cationic surfactants benzethonium  
368 and benzalkonium. Most of these genes are involved in membrane lipid homeostasis,  
369 envelope stress response pathways and drug efflux systems (**Fig. 4, Supplementary**  
370 **Table 3**).

371  
372 In total, we identified 41 instances where the Dub-seq fitness data is consistent with the  
373 known growth benefit imparted by the gene (**Supplementary Table 4**). These high  
374 confidence, known hits include genes encoding diverse functions such as efflux pumps,  
375 transporters, and regulators, as well as biosynthetic enzymes and small RNAs, each  
376 yielding enhanced fitness via diverse mechanisms. For example, overexpression of  
377 *cysE* (which encodes serine acetyltransferase) probably increases nickel tolerance  
378 through increased glutathione biosynthesis<sup>53</sup>, while overexpression of *rnc* (which  
379 encodes RNase III) yields a growth benefit in nickel and cobalt stress, as it down-  
380 regulates the expression of *corA*, which encodes a transporter that mediates the influx  
381 of nickel and cobalt ions into the cell<sup>54</sup>.

382  
383 In addition to the known cases, we also identified hundreds of genes that had not been  
384 previously associated with a tolerance phenotype in a specific condition, including *pssA*,  
385 *dcrA/sdaC*, *dcrB* in sisomicin; *pmrD* in aluminum; *treA*, *treB* and *phnM* in phosphomycin;  
386 sRNAs *chiX* in nickel and *ryhB* in zinc; and many genes of unknown function (**Fig. 4,**  
387 **Supplementary Table 3**). To follow up some of the novel observations, we assayed the  
388 growth of strains overexpressing the genes individually with and without added stress.  
389 We used *murA* overexpression as a test case, as this is known to confer resistance to  
390 phosphomycin<sup>55</sup> (**Supplementary Fig. 5**). Growth curves confirmed that the

391 overexpression of either *pssA* or *dcrB* confers resistance to the aminoglycoside  
392 antibiotic sisomicin, although the mechanism(s) by which this resistance is conferred  
393 remains unclear. The gene *pssA* encodes an essential phosphatidylserine synthase,  
394 while *dcrB* is a periplasmic protein with a role in phage infection<sup>48</sup>. Growth curves also  
395 confirm that the overexpression of the outer membrane protein MipA confers strong  
396 resistance to benzethonium chloride (**Supplementary Fig. 5**). *mipA* has previously  
397 been implicated in the resistance to other antibiotics<sup>56</sup>.

398  
399 Gene overexpression can also decrease host fitness<sup>16,17,46</sup> and may indicate important  
400 function for those gene products. We identified 570 genes with a high-confidence  
401 negative effect on fitness in at least one experiment (**Supplementary Table 3**). Some of  
402 these genes appear to be more generally toxic when overexpressed or have a global  
403 regulatory role and compromise host fitness in multiple conditions. 24 genes had  
404 detrimental effects on fitness in 10 or more different conditions (*ampH*, *arcZ*, *aroK*, *crr*,  
405 *gadY*, *hfq*, *hha*, *htpX*, *hupB*, *iraP*, *metJ*, *mtlA*, *nupG*, *rpoS*, *ruvA*, *tsx*, *wecA*, *ybjT*, *yceG*,  
406 *ydgA*, *ydjN*, *yibN*, *yjdC*, and *zinT*). Conversely, some genes have negative gene scores  
407 in only one or a handful of conditions. For example, consistent with earlier studies we  
408 found that overexpression of *glpT* or *uhpT* increases susceptibility to phosphomycin<sup>57</sup>.  
409 These results also agree with clinical data, which shows that the main cause of  
410 phosphomycin resistance in patients is the down-regulation of GlpT via down-regulation  
411 of cAMP<sup>57</sup>. Accordingly, we also found that overexpression of *cpdA* (which encodes an  
412 enzyme that hydrolyzes cAMP) enhances fitness under phosphomycin stress (**Fig. 4**).

413  
414 Finally, we analyzed our data for ‘epistatic’ instances where multiple genes on a  
415 fragment are necessary for the observed phenotype. Specifically, we searched for  
416 evidence of synergy between genes by analyzing scores for fragments containing more  
417 than one gene that are significantly greater than the inferred sum of score of the  
418 constituent genes (Methods). In total, we found 6 high scoring epistatic-effect cases  
419 across 52 conditions in our Dub-seq dataset (*fetA-fetB* on nickel, *ampD-ampE* on  
420 benzethonium, *ackA-pta* on D-lactate, *arcA-yjiY* on sisomicin, *hns-tdk* on phosphomycin  
421 and *yfiF-trxC* on potassium acetate (**Supplementary Fig.6abc**)). Among these, 3 gene-  
422 pairs have related functions (*fetA-fetB* form a complex, *pta-ackA* encode enzymes that  
423 catalyze adjacent reactions in the catabolism of lactate, and *ampD-ampE* are thought to  
424 be a signaling pathway<sup>48</sup>) and our data indicates, together they provide a larger growth  
425 benefit. Specifically, overexpression of *fetAB* together has been shown to improve  
426 survival during nickel stress<sup>58</sup>.

#### 427 428 **Comparison to loss-of-function fitness data**

429 Integrating large-scale genetic gain and loss of function can provide added specificity to  
430 biological insights. For instance, genes with resistance phenotypes when

431 overexpressed and sensitivity phenotypes when deleted are often specifically involved  
432 in the condition of interest, as demonstrated by studies identifying drug targets in  
433 yeast<sup>59</sup> or identifying small RNA regulators<sup>60</sup> or antibiotic resistance factors in bacteria<sup>61</sup>.  
434 Furthermore, genes with opposing loss and gain-of-function phenotypes for stress  
435 compounds are more likely to be true resistance determinants as opposed to genes that  
436 have indirect effects when overexpressed<sup>16</sup>. For 45 of the conditions that we profiled in  
437 this study with Dub-seq, we can systematically compare these phenotypic  
438 consequences of overexpression to loss-of-function mutations as determined by  
439 random barcode transposon site mutagenesis<sup>15</sup>. The two data sets studied the same  
440 growth media and compounds, but not necessarily at the same concentrations, and they  
441 used different strains of *E. coli* (DH10B or BW25113). Across these 45 conditions, we  
442 identified 625 high-confidence benefits of overexpression (or 0.3% of gene-condition  
443 pairs). Of the 625 high-confidence benefits, 480 are for genes with RB-TnSeq data, and  
444 in 62 cases (12%), that loss of function led to a significant disadvantage (RB-TnSeq  
445 fitness < -1 and  $t < -4$ , where  $t$  is a t-like test statistic<sup>13</sup>). By chance, we would expect  
446 just 2.5% agreement, which is significantly less ( $P < 10^{-15}$ , chi-squared test of  
447 proportions). Overall, we found moderate overlap between genes that are beneficial  
448 when overexpressed and important for fitness when disrupted (**Supplementary Table**  
449 **3**).

450  
451 To illustrate the biological insights that can be derived by systematically comparing gain  
452 and loss-of-function data on a genomic scale, we present 3 examples: growth in the  
453 presence of elevated nickel, cobalt, or sodium chloride (**Fig. 5abc**). Under each  
454 condition, we find that a number of genes that are both necessary for resisting the  
455 stress when knocked-out and sufficient for a resistance phenotype when singly  
456 overexpressed. These instances include known examples such as the aforementioned  
457 metal exporter RcnA<sup>45</sup> and RNase III for cobalt and nickel tolerance<sup>54</sup>, as well as the  
458 osmolyte transporter ProP<sup>62</sup> and envelope biogenesis factor YcbC (ElyC)<sup>63</sup> for tolerance  
459 to osmotic stress imposed by sodium chloride. (In our Dub-seq data, *proP* and *ycbC*  
460 failed to pass the filters for high-confidence effects). In addition to these known  
461 examples, there are more novel observations (**Fig. 5abc**). Under nickel and cobalt  
462 stress, the uncharacterized protein YfgG (DUF2633) is important for tolerance, a finding  
463 that is supported by RB-Tnseq data<sup>15</sup> and by individual growth curve analysis of an *yfgG*  
464 overexpression strain (**Fig. 5d**). While the precise biochemical function of YfgG is  
465 unclear, a close homolog of this protein in *Klebsiella michiganensis* is also important for  
466 fitness under nickel and cobalt stress<sup>15</sup>. As a second example, we find that ProY is  
467 important for nickel resistance. A ProY homolog in the related bacterium *K.*  
468 *michiganensis* is also important for nickel resistance<sup>15</sup>. Using individual strain growth  
469 curve analysis, we confirmed that overexpression of *proY* alone can confer nickel  
470 resistance to *E. coli* (**Fig. 5e**). While ProY is currently annotated as a cryptic proline

471 transporter, we suspect that its function is to transport histidine as it can suppress  
472 histidine auxotrophy<sup>25</sup> and homologs of this protein are required for histidine utilization  
473 in other bacteria<sup>15</sup>. In light of this, we speculate that the nickel resistance phenotype of  
474 ProY is due to increased sequestration of nickel ions by a higher intracellular  
475 concentration of histidine. As a final example, we found that the porphyrogen oxidase  
476 YfeX confers sodium chloride resistance in *E. coli*, a finding confirmed by an individual  
477 growth curve analysis (**Fig. 5f**). While we are unsure how this protein manifests this  
478 phenotype, we note that yfeX homologs are important for resisting sodium chloride in  
479 multiple bacteria<sup>15</sup>. We have provided a general working hypothesis for many of other  
480 genes with high fitness scores in **Supplementary Table 5**.

481

## 482 **DISCUSSION**

483

484 Here we describe Dub-seq, a technology for performing parallelized gain-of-function  
485 fitness assays across diverse conditions. Dub-seq couples shotgun cloning of random  
486 DNA fragments with competitive fitness assays to assess the phenotypic importance of  
487 the genes contained on those fragments in a single tube assay. We demonstrate that  
488 Dub-seq is reproducible, economical, scalable, and identifies both known and novel  
489 gain-of-function phenotypes. By decoupling the library creation and characterization  
490 step from the screening step with BarSeq, Dub-seq provides a quantitative and rapid  
491 tool for experimentally assessing gene function via overexpression phenotypes of DNA  
492 cloned into an expression vector. This approach can improve overall repeatability and  
493 reproducibility of genome-wide gain-of-function experiments, and facilitate open  
494 distribution of libraries among researchers<sup>64</sup>.

495

496 In this proof-of-concept study, we generated a Dub-seq library of *E. coli* genomic DNA  
497 in a broad-range expression vector and assayed the phenotypic importance of  
498 overexpressing cloned genes using *E. coli* as the host bacterium. From 152 genome-  
499 wide assays, we identified 400 different genes with a high-confidence fitness benefit  
500 when overexpressed in at least one experimental condition. The majority of these gene-  
501 phenotype associations have not previously been reported including, as far as we know,  
502 for *yfgG*, *proY*, and *yfeX* (**Supplementary Table 3**). We found 241 genes confer a  
503 fitness benefit in just one condition, indicating a condition-specific phenotype. Overall,  
504 32 genes enhanced fitness in 5 or more conditions, suggesting their broader role in host  
505 fitness and importance in cross-resistance phenotypes observed between metals,  
506 antibiotics, antiseptics and other stresses<sup>65</sup>. Dub-seq recapitulated 41 known instances  
507 of positive fitness effects, wherein the fitness phenotypes stem from diverse  
508 mechanisms, including overexpression of a compound target, active efflux of heavy  
509 metals, decreased uptake of metals and antibiotics, increased uptake of nutrients, and  
510 the regulatory effects of both protein-coding genes and small RNAs. We also identified

511 enhanced susceptibility due to overexpression. Finally, we show that systematically  
512 comparing gain and loss-of-function datasets provide additional insights into those  
513 genes that are both necessary and sufficient for stress tolerance phenotypes.

514  
515 Dub-seq can be readily extended to DNA from other sources and many cultured  
516 bacteria could be adapted as hosts for the genome-wide fitness assays. In particular,  
517 our vectors should be suitable to build Dub-seq libraries of microbial isolates and can be  
518 mobilized to new bacteria via conjugation because of its broad-host range replication  
519 origin. By using other hosts, we can overcome gene expression and toxicity issues  
520 associated with expressing heterologous DNA in model hosts<sup>34-36</sup>. To extend the Dub-  
521 seq methodology for functional profiling of DNA isolated from the environment, we  
522 would need to generate a higher diversity of barcoded vectors so that we would have a  
523 large library of unique barcode pairs and the largest percentage of metagenomic  
524 diversity can be captured and mapped confidently. In addition, to ensure reliable  
525 expression of heterologous genes, a number of approaches can be used to activate  
526 transcription or translation of genes encoded within foreign DNA<sup>34,42,66</sup>.

527  
528 In this work, we generated a Dub-seq library with a ~2.6 kb insert size and therefore by  
529 design, the library only covers fragments encoding 2-3 genes on an average. Therefore,  
530 phenotypes that are only conferred by the activity of a larger group of genes (such as  
531 multisubunit complexes) will not be detected. Nevertheless, we did detect 6 instances of  
532 'epistatic' interactions in which two neighboring genes show greater fitness score as  
533 gene-pairs than the inferred sum of score of the individual genes. By adapting the Dub-  
534 seq strategy to fosmids, cosmids and bacterial-artificial-chromosomes, future efforts can  
535 clone larger size genomic fragments to create Dub-seq libraries for the discovery of  
536 activities encoded by multiple genes, including secondary metabolites.

537  
538 Given the increasing knowledge gap between genomic sequence and function, and the  
539 limited ability of computational approaches to accurately predict gene function from  
540 sequence, high-throughput experimental methods are needed to assign gene function  
541 and resolve roles of uncharacterized genes. Recently, a number of loss-of-function  
542 methods have been developed<sup>5-8,10-14</sup>, but only a fraction of genes from genetically  
543 tractable microbes can be readily annotated with a specific function using these  
544 approaches. We envision that multiple, complementary experimental approaches that  
545 can be applied *en masse* are ultimately necessary to uncover the roles of most poorly  
546 annotated genes from microbial isolates and microbiomes. The Dub-seq approach we  
547 presented here is another valuable tool in this toolkit.

548  
549  
550

551 **Author contributions**

552 V.K.M., A.M.D. and A.P.A. conceived the project. V.K.M., A.M.D., A.P.A., supervised  
553 the project. V.K.M. led the experimental work. P.S.N. led the computational work.  
554 V.K.M., A.M.D., T.K.O., M.C. and S.C. collected data. V.K.M., P.S.N., M.N.P. and  
555 A.M.D. analyzed the fitness data. M.N.P. and A.P.A. provided advice on data  
556 processing and modeling. V.K.M., P.S.N., M.N.P., A.M.D. and A.P.A. wrote the paper.

557

558 **Acknowledgements**

559 We thank Mahek Modi and Aaron Gupta for assisting in the initial stage of this project.  
560 The initial concepts for this project were developed by Biodesign project supported by  
561 the Office of Science (BER), U.S. Department of Energy, DE-SC0008812. The  
562 implementation was funded by ENIGMA, a Scientific Focus Area Program at Lawrence  
563 Berkeley National Laboratory, supported by the U.S. Department of Energy, Office of  
564 Science, Office of Biological and Environmental Research under contract DE-AC02-  
565 05CH11231.

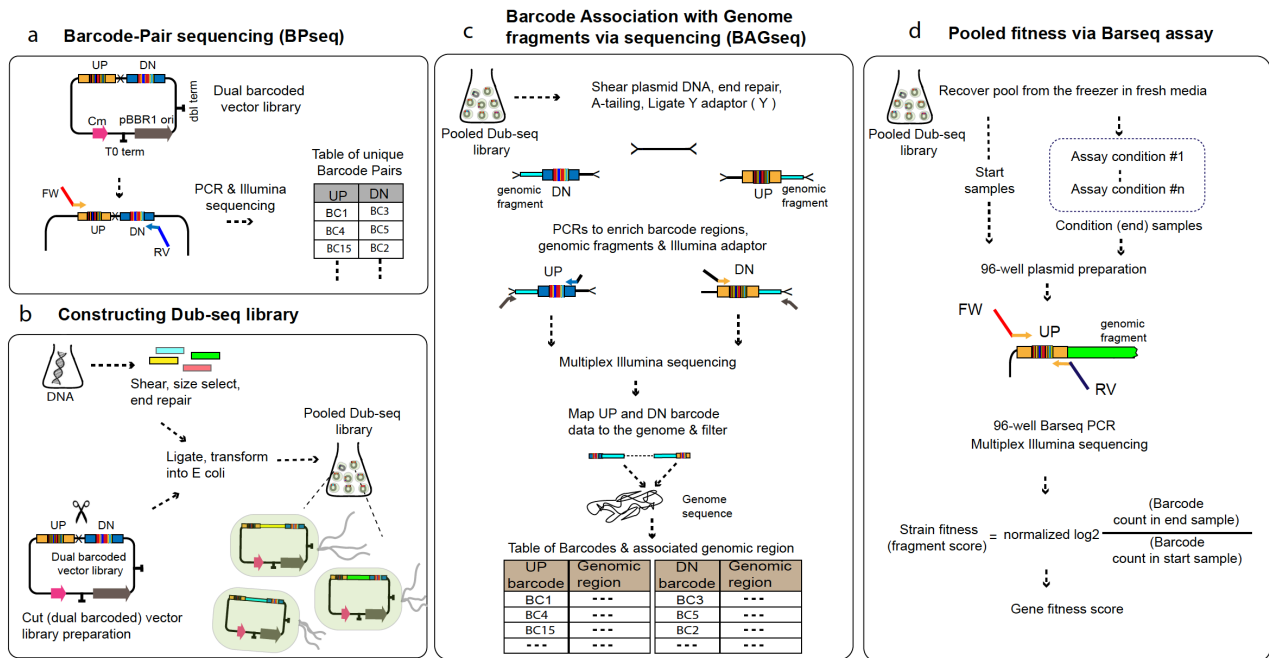
566

567 **Competing interest**

568 VKM, PSN, AMD, and APA are holders of a patent on the Dub-seq technology.

569

570  
571 **FIGURES**  
572  
573

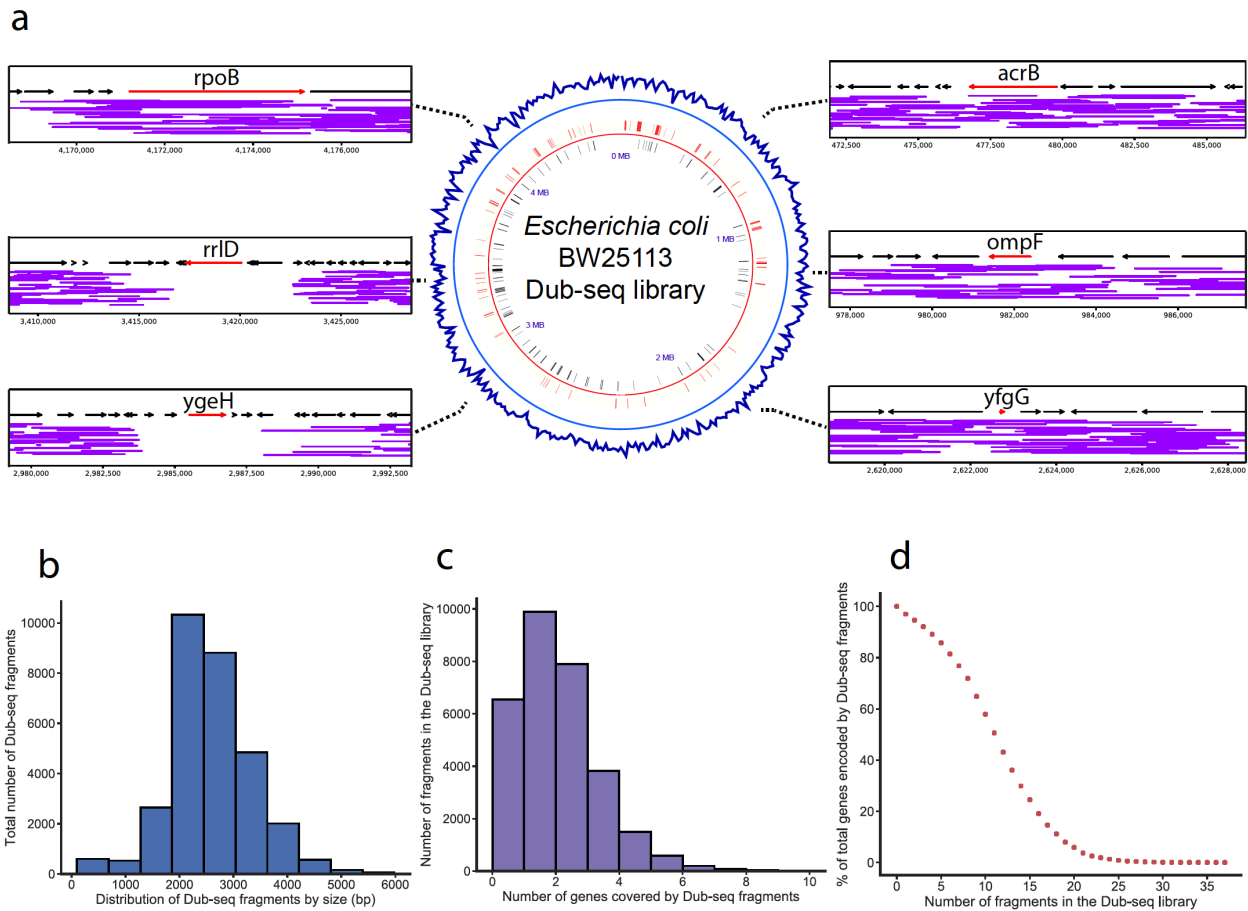


574  
575  
576 **Figure 1. Schematic overview of the Dub-seq approach.** (a) A pair of random 20  
577 nucleotide DNA sequences, the UP and DOWN (DN) barcodes are cloned into an  
578 expression vector. Deep sequencing of the dual barcoded vector (BPseq) associates  
579 UP and DOWN barcode sequences. (b) Target genomic DNA is randomly sheared and  
580 cloned between the UP and DOWN barcodes to create the Dub-seq plasmid library. (c)  
581 To characterize the Dub-seq library, a “Tn-seq” like protocol is performed to precisely  
582 map the two genomic breakpoints of each insert and to associate each breakpoint with  
583 its random DNA barcode sequence. If the source genome(s) has been sequenced, then  
584 BAGseq can be used to define the exact sequence of each plasmid in the library. (d)  
585 The fitness of bacteria carrying different plasmids can be measured with pooled growth  
586 assays and deep sequencing of the DNA barcodes (BarSeq). Strain (or fragment)  
587 fitness is defined as the  $\log_2$  ratio of barcode abundance after selection (end) versus  
588 before (start). Gene fitness is estimated from the fragments’ fitness by a constrained  
589 regression.

590



591



592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

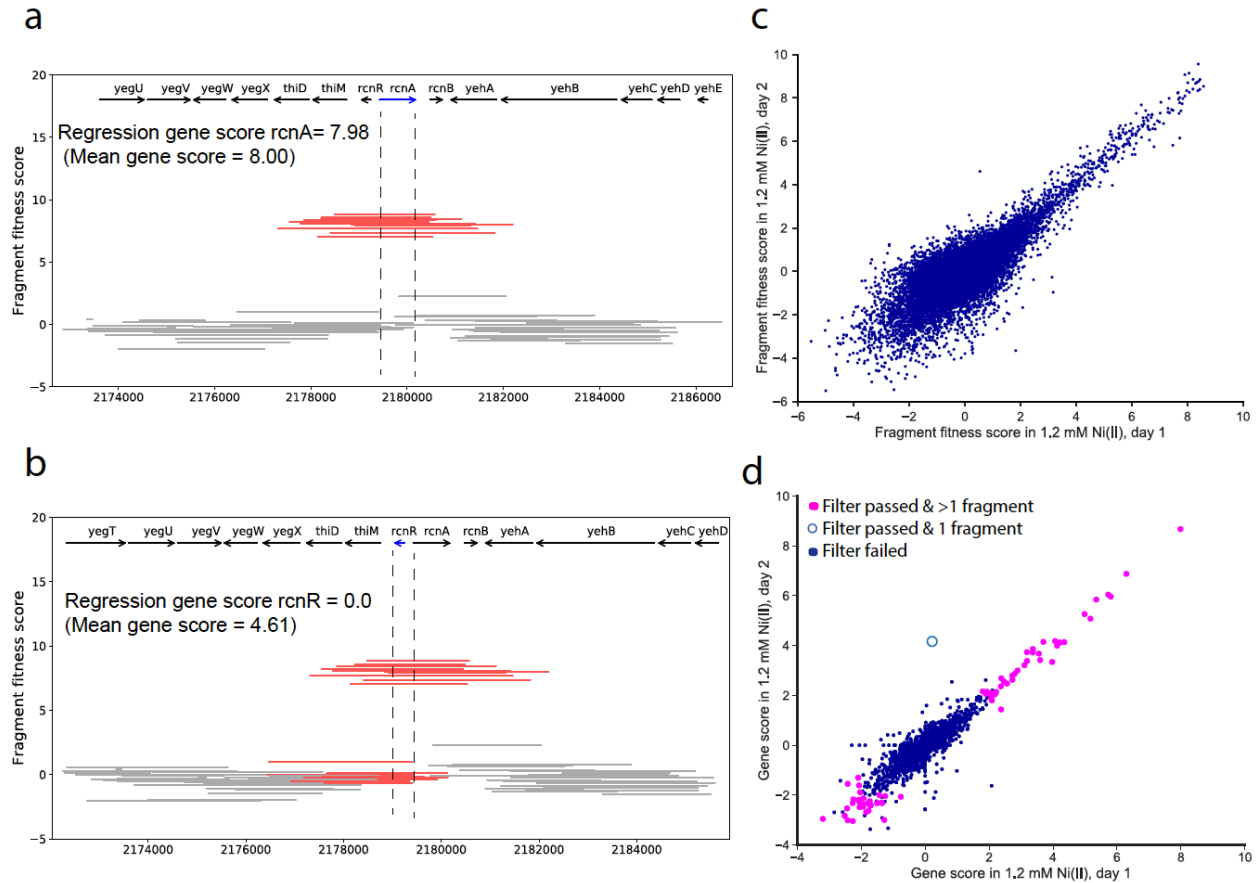
607

608

609

610

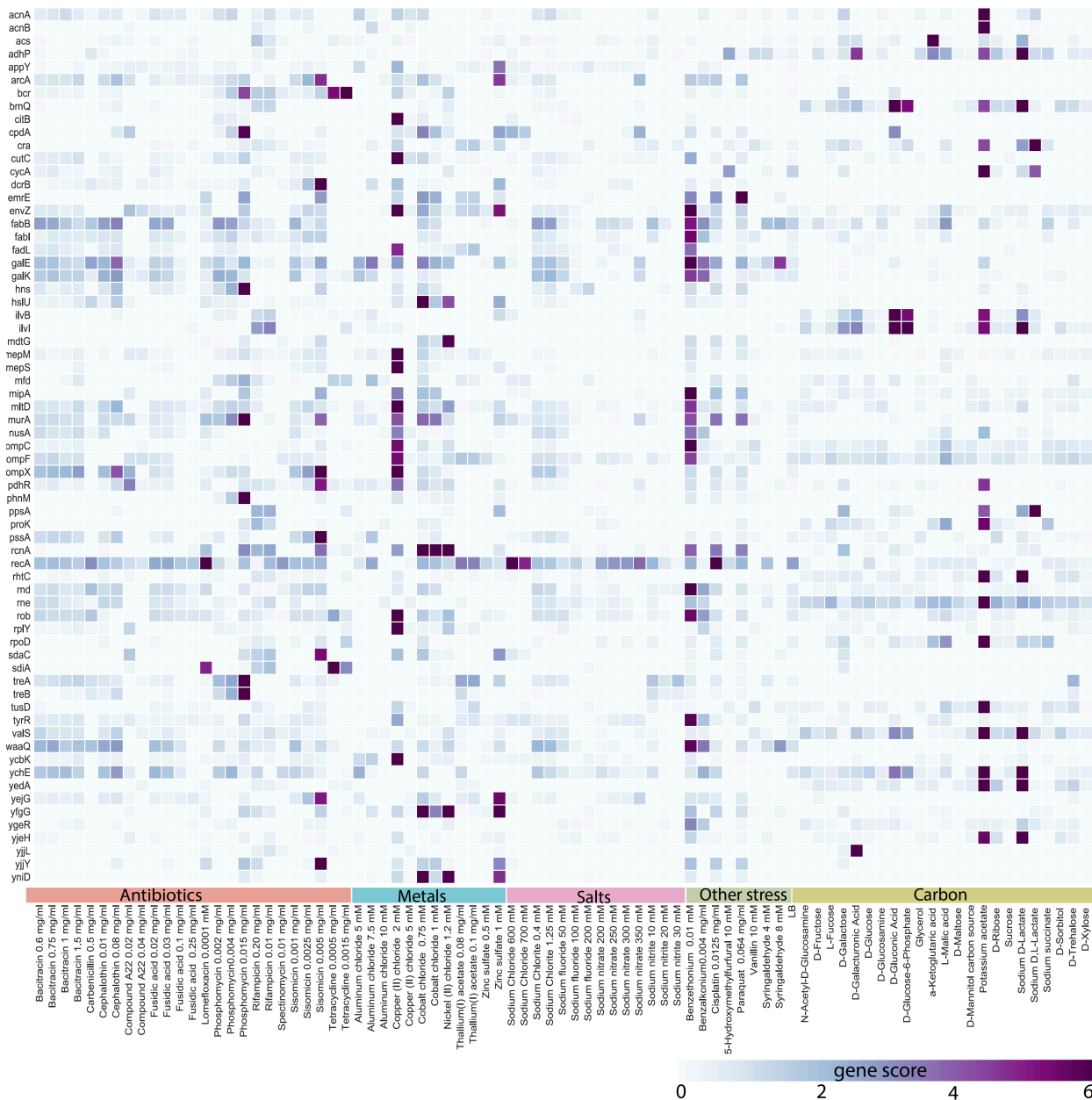
**Figure 2. *E. coli* Dub-seq library characterization.** (a) Center: genomic coverage of the *E. coli* BW25113 Dub-seq library in 10 kB windows (blue track). Black and red line-tracks represent genes essential for viability when deleted<sup>5</sup> that are encoded on the negative and positive strands, respectively and are covered in the Dub-seq library. Left and right: regions of the *E. coli* chromosome covering *acrB*, *ompF*, *yfgG*, *ygeH*, *rrlD* and *rpoB*. Each purple line represents a Dub-seq genomic fragment (the y-axis is random). (b) The fragment insert size distribution in the *E. coli* Dub-seq library. (c) The distribution of number of genes that are completely covered (start to stop codon) per genomic fragment in the *E. coli* Dub-seq library. (d) Cumulative distribution plot showing the percentage of genes in the *E. coli* genome (y-axis) covered by a number of independent genomic fragments (x-axis).



611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623

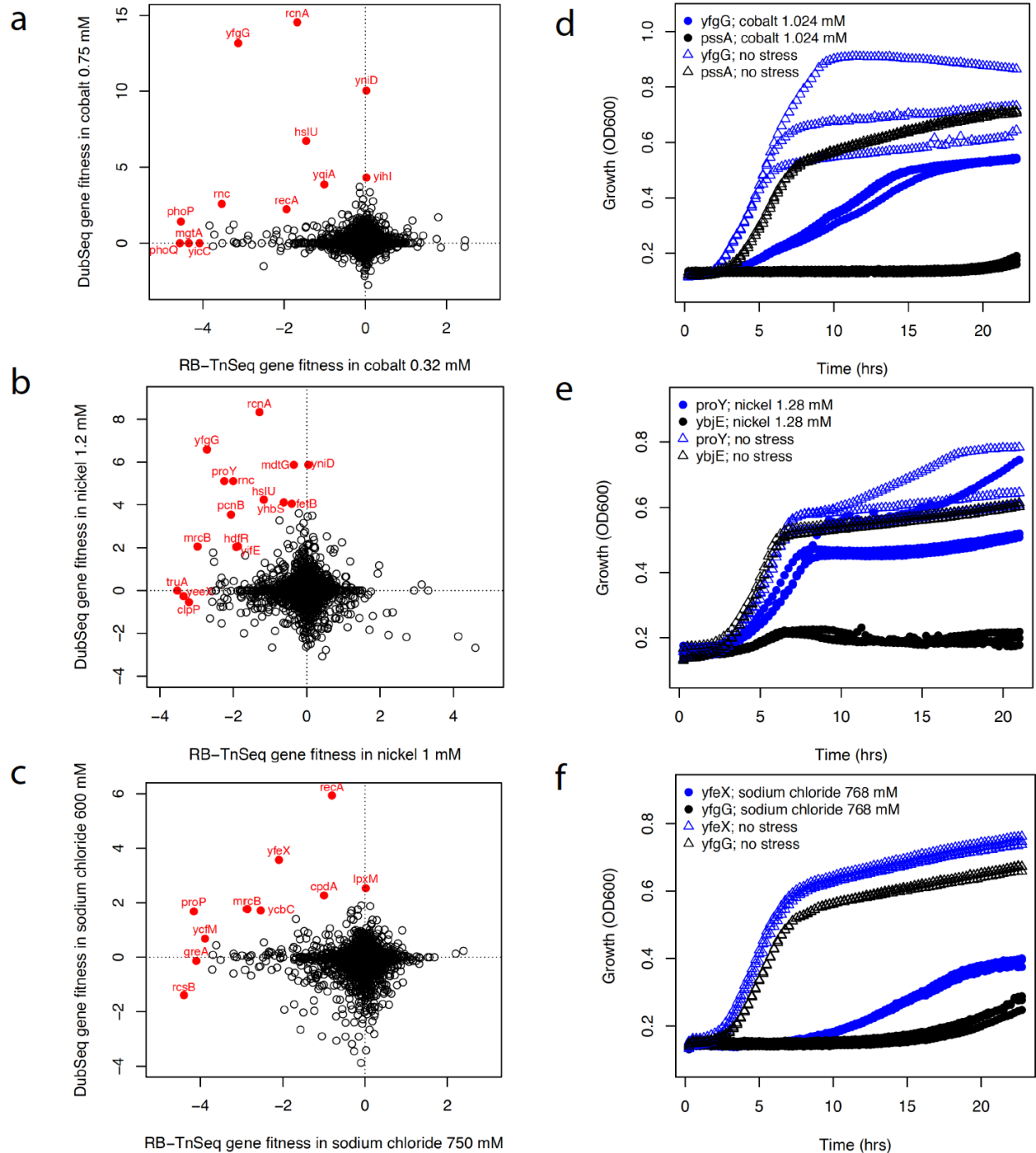
**Figure 3. Fragment and gene fitness Dub-seq scores.** (a) Dub-seq fragment (strain) data for region surrounding *rcnA* under elevated nickel stress (y-axis). Each line shows a Dub-seq fragment. Those that completely cover *rcnA* are in red. Both the mean and regression scores reflect the known biology of *rcnA* as a nickel resistance determinant<sup>45</sup>. (b) Same as (a) for the neighboring *rcnR*, which encodes a transcriptional repressor of *rcnA*. Fragments that cover *rcnR* are in red. (c) Comparison of fragment fitness scores for two biological replicates of 1.2 mM nickel stress. (d) Same as (c) for gene fitness scores calculated using the regression approach. Genes are highlighted if their data passed our statistical filters for reliable effects (see Methods); we also show whether the gene score is based on just one fragment.

624  
625  
626



627  
628  
629  
630  
631  
632  
633  
634

**Figure 4. Heatmap of Dub-seq fitness data for 53 conditions and for 67 genes with large benefits.** Only genes with a high-confidence effect and gene fitness score  $\geq 6$  in at least one condition are shown. Gene scores from replicate experiments were averaged.



635  
636

637 **Figure 5. Comparing genome-wide loss and gain-of-function phenotype**  
638 **data.** Comparison of RB-TnSeq fitness data<sup>15</sup> (x-axis) and Dub-seq gene fitness data  
639 for *E. coli* genes under growth with inhibitory concentrations of cobalt (a), nickel (b), and  
640 sodium chloride (c). Selected genes are highlighted. (d) Growth of *E. coli*  
641 overexpressing *yfgG* under cobalt stress; *pssA* is a control. (e) Growth of *E. coli*  
642 overexpressing *proY* under nickel stress; *ybjE* is a control. (f) Growth of *E. coli*  
643 overexpressing *yfeX* under sodium chloride stress; *yfgG* is used as a control.

644 **METHODS:**

645

646 **Strains and growth conditions**

647 *Escherichia coli* BW25113 was purchased from the *E. coli* Genetic Stock Center. All  
648 plasmid manipulations were performed using standard molecular biology techniques<sup>67</sup>.  
649 All enzymes were obtained from New England Biolabs (NEB) and oligonucleotides were  
650 received from Integrated DNA Technologies (IDT). *Escherichia coli* strain DH10B  
651 (DH10B derivative, NEB 10-Beta) was used for plasmid construction and as host for  
652 Dub-seq fitness assays. Unless noted, all strains were grown in LB supplemented with  
653 30 µg/ml chloramphenicol at 37°C and shaking at 200 rpm. The primers, plasmids and  
654 strains used in this study are listed in **Supplementary Tables 6, 7 and 8** respectively.

655

656 **Construction of dual barcoded Dub-seq vector**

657 To construct a double barcoded vector, we used pFAB5477 an in-house plasmid with  
658 pBBR1 replication origin and a chloramphenicol resistance marker<sup>68</sup>. pBBR1 based  
659 broad-host plasmids are relatively small, mobilizable and have been widely used for a  
660 variety of genetic engineering applications in diverse microbes<sup>69</sup>. To insert a pair of DNA  
661 barcodes on the plasmid we used phosphorylated oFAB2853 and oFAB2854 primers to  
662 amplify the entire plasmid pFAB5477, removed the plasmid backbone using DpnI (as  
663 per manufacturing instructions, NEB), and ligated the amplified and pure product using  
664 T4 ligase (as per manufacturing instructions, NEB). The random N's in oFAB2853 and  
665 oFAB2854 (**Supplementary Table 6**) represent the UP and DOWN barcode  
666 sequences. The ligated product, pFAB5491, was column purified using the Qiagen PCR  
667 purification kit, transformed into DH10B electro-competent cells (NEB 10-Beta *E. coli*  
668 cells, as per manufacturing instructions, NEB) and transformants were selected on LB-  
669 agar plates supplemented with 30 µg/ml chloramphenicol. The next day, ~250,000  
670 colony forming units (CFU) were estimated and scraped together into 20 ml LB with 30  
671 µg/ml chloramphenicol. The culture library was diluted to an optical density at 600 nm  
672 (OD600) of 0.2 in fresh LB medium supplemented with 30 µg/ml chloramphenicol and  
673 grown to a final OD600 of ~1.2. We added glycerol to a final concentration of 15%,  
674 made multiple 1 ml glycerol stocks, and stored them at -80°C. We also collected cell  
675 pellets to prepare plasmid DNA of pFAB5491 for further characterization of the library  
676 (BPseq).

677

678 **BPseq to characterize dual barcoded Dub-seq vector**

679 To associate the pair of DNA barcodes, we performed Barcode-Pair sequencing  
680 (BPseq) of the plasmid pFAB5491 library. For deep coverage of the library, we  
681 performed 10 different PCR reactions using primers VM\_barseq\_P1 and VM\_Barseq-  
682 P2. The forward primers VM\_Barseq-P2 contains different 6-bp TruSeq indexes, and  
683 were automatically demultiplexed by the Illumina software.

684  
685 We performed PCR in a 100- $\mu$ l total volume with 5  $\mu$ l common reverse primer  
686 VM\_barseq\_P1 (4  $\mu$ M), 5  $\mu$ l forward primer VM\_Barseq-P2\_IT001 to IT010 (4  $\mu$ M), 38  
687  $\mu$ l of sterile water, 2  $\mu$ l template pFAB5491, and 50  $\mu$ l of 2X stock of Q5 DNA  
688 Polymerase mix (500  $\mu$ l of 2X stock of Q5 DNA Polymerase mix consists of 200  $\mu$ l Q5  
689 buffer, 20  $\mu$ l dNTP, 50  $\mu$ l DMSO, 10  $\mu$ l Q5 DNA Polymerase enzyme and 220  $\mu$ l water)  
690 under following PCR conditions: 98°C for 4 minutes, followed by 15 cycles of 30 sec at  
691 98°C, 30 sec at 55°C, 30 sec at 72°C and final extension at 72°C for 5 minutes. Finally,  
692 we ran the PCR products on an analytical gel to confirm amplification. We pooled equal  
693 volumes (10  $\mu$ l) of BarSeq PCR products, purified the combined product using Qiagen  
694 PCR purification kit, and eluted in 40  $\mu$ l of sterile water. We quantified the DNA product  
695 with a Qubit double-stranded DNA (dsDNA) high-sensitivity (HS) assay kit (Invitrogen).  
696 The BPseq samples were sequenced first on Illumina MiSeq and then HiSeq 2500: both  
697 with 150 bp single-end runs.

### 698 699 **BPseq data analysis**

700 BPseq reads were analyzed with *bpseq* script from the *Dub-seq* python library with  
701 default parameters (code available at <https://github.com/psnovichkov/DubSeq>). The  
702 script looks for the common flanking sequences around each barcode (UP and DOWN)  
703 and requires an exact match of 9 nucleotides on both sides. By default, these flanking  
704 sequences may be up to 2 nucleotides away from their expected positions. The script  
705 also requires that each position in each barcode have a quality score of at least 20 (that  
706 is, an estimated error rate of under 1%). This gives an initial list of pairs of barcodes  
707 with the correct length and reliable sequence quality.

708  
709 We applied two additional filters to minimize the number of erroneous barcode pairs that  
710 can be caused by PCR artifacts or sequencing errors. First, we check whether a given  
711 barcode can be a result of a single nucleotide substitution introduced in a real barcode  
712 and filter out all such barcodes. We perform a pairwise sequence comparison of all  
713 extracted barcodes (UP and DOWN barcodes are treated separately) and search for  
714 “*similar*” barcodes. Two barcodes are considered to be *similar* if they are different by  
715 only one nucleotide. A given barcode passes the filter if it does not have similar  
716 barcodes or it is at least two times more frequent than the most abundant similar  
717 barcode.

718  
719 Second, we check whether a given barcode pair can be a result of chimeric PCR and  
720 filter out all such pairs. As the region between and around UP and DOWN barcodes are  
721 identical in all plasmids in our library, we expected artifacts from formation of chimeric  
722 BPseq PCR products<sup>13</sup>. We perform a pairwise comparison of all barcode pairs and  
723 search for “*related*” pairs. Two barcode pairs are considered to be related if they have

724 either the same UP or DOWN barcodes. The presence of the same UP (or DOWN)  
725 barcode in multiple barcode pairs is potentially a sign of chimeric PCR. To distinguish  
726 the true barcode pair from the chimeric one, we check the frequency of all the related  
727 barcode pairs. A given barcode pair passes the filter and is considered to be non-  
728 chimeric if it does not have related pairs or it is at least two times more frequent than the  
729 most abundant related barcode pair. As a result, the 'reference set' of barcode pairs is  
730 created. From the BPseq step we obtained 5,436,798 total reads. Among these, total  
731 usable reads (reads that support barcode pairs from the reference set) were 2,933,702  
732 and represent about 54% of total reads.

733

### 734 **Dub-seq vector preparation for cloning genomic fragments**

735 To prepare the Dub-seq vector pFAB5491 for cloning, we made 900 ul or about 100 ug  
736 of plasmid preparation (Qiagen plasmid miniprep kit), and performed two rounds of Pml  
737 digestion. Restriction digestion reaction included 900 ul (total 100 ug) of pFAB5491  
738 plasmid, 100 ul Pml enzyme, 400 ul 10X cutsmart buffer, and water to make up the  
739 volume of 4000 ul. We incubated the reaction at 37°C on a heating block for 4 hours  
740 and then checked the reaction progress on an analytical 1% agarose gel. To  
741 dephosphorylate the restriction-digested vector, we added 1 unit of rSAP for every 1  
742 pmol of DNA ends (about 1 µg of a 3 kb plasmid), and incubated at 37°C for 2 hours in  
743 a PCR machine. We stopped the reaction by heat-inactivation of rSAP and restriction  
744 enzyme at 70°C for 20 minutes. The cut and dephosphorylated vector library was then  
745 gel purified (Qiagen gel extraction kit). To remove any uncut vector, we repeated the  
746 entire process of restriction digestion, dephosphorylation, and purification. The final  
747 concentration of cut and pure barcoded vector library used for cloning genome  
748 fragments was about ~30 ng/ul.

749

### 750 **Construction of *E. coli* Dub-seq library**

751 To construct Dub-seq library of *E. coli* genomic fragments, we extracted *E. coli*  
752 BW25113 genomic DNA and 1 ug was fragmented by ultrasonication to an average size  
753 of 3000 bp with a Covaris S220 focused ultrasonicator. The sheared genomic DNA was  
754 then gel purified and end-repaired using End-IT kit (Epicentre, as per manufacturer  
755 instruction). Briefly the 50 ul reaction included: 34 ul sheared DNA (1.0 ug total), 5 ul  
756 ATP 10 mM, 5 ul dNTP mix (10 mM), 5 ul EndIt buffer 10X and 1-2 ul EndIT enzyme.  
757 We incubated the reaction at room temperature for 45 mins, and inactivated the enzyme  
758 by incubating the reaction at 70°C for 10 minutes. The end-repaired genome fragments  
759 were purified with PCR clean-up kit (Qiagen), and quantified on Nanodrop.

760

761 The end-repaired genomic fragments were then ligated to the restriction-digested,  
762 sequence-characterized dual barcoded backbone vector (pFAB5491) at 8:1  
763 insert:vector ratio using Fast-link Ligase enzyme (Epicentre, as per manufacturer

764 instruction). The total 60 ul ligation reaction consists of 4 ul of restriction-digested  
765 pFAB5491, 20 ul End-repaired DNA, 3 ul ATP (10 mM), 6 ul 10X ligase buffer, 19 ul  
766 water and 8 ul Fast-link-ligase. The ligation was incubated overnight (18 hrs) at 16°C,  
767 inactivated at 75°C for 15 minutes, and purified using PCR purification kit (Qiagen).

768  
769 For transforming the ligation reaction, 60 ul of column-purified ligation reaction was  
770 mixed gently with 1500 ul of NEB DH10B electrocompetent cells on ice and then the  
771 mix was dispensed 60 ul per cuvette. Electroporation was done using parameters  
772 supplied by NEB. Transformed cells were recovered by adding 1 ml SOC recovery  
773 media (as per competent cell manufacturer instruction, NEB). We pooled all recoveries  
774 and added additional 10 ml of fresh SOC. Transformants were then incubated at 37°C  
775 with shaking for 90 minutes. We spun down the pellets and resuspended the pellet in 6  
776 ml SOC. Different volumes of 6 ml resuspended pellets were then plated on overnight-  
777 dried bioassay plates (Thermo Scientific # 240835) of LB agar supplemented with 30  
778 ug/ml chloramphenicol. We also did dilution series for estimating CFUs.

779  
780 We determined the number of colonies required for 99% coverage of *E. coli* genome  
781 using the formula  $N = \ln(1-0.99)/\ln(1-(\text{Insert size}/\text{Genome Size}))$  to ensure that genome  
782 fragments are present in the cloned library<sup>70</sup>. For example, to cover the *E. coli* genome  
783 (of size 4.7 Mb) with fragments of 3 kb, we need about 4,610 strains for 99% coverage.  
784 We collected ~40,000 colonies by scraping the colonies using a sterile spatula into 20  
785 ml LB supplemented with 30 ug/ml chloramphenicol in a 50 ml Falcon tube and mixed  
786 well. This *E. coli* Dub-seq library was then diluted to an optical density at 600 nm  
787 (OD600) of 0.2 in fresh LB supplemented with 30 ug/ml chloramphenicol and grown to a  
788 final OD600 of ~1.2 at 37°C. We added glycerol to a final concentration of 15%, made  
789 multiple stocks of 1 ml volume, and stored the aliquots at -80°C. We also made cell  
790 pellets to store at -80°C and to make large plasmid preparation (Qiagen) for BAGseq  
791 library preparation.

792

### 793 **BAGseq to characterize barcoded genomic fragment junctions**

794 We characterized the final plasmid library pFAB5516 using a TnSeq-like protocol<sup>13</sup>,  
795 which we call Barcode-Association-with Genome fragment sequencing or BAGseq.  
796 BAGseq identifies the cloned genome fragment and its pairings with neighboring dual  
797 barcodes. This step of associating the dual barcodes with each library of genomic  
798 fragments is only done once (by deep sequencing) and used as a reference table to  
799 derive connections between observed functional/fitness traits with specific cloned  
800 genomic fragment (Fig. 1).

801  
802 To generate Illumina-compatible sequencing libraries to link both UP and DOWN  
803 random DNA barcodes to the ends of the cloned genome fragments, we processed two



804 samples per library. The plasmid library (1 ug) samples were fragmented by  
805 ultrasonication to an average size of 300 bp with a Covaris S220 focused ultrasonicator.  
806 To remove DNA fragments of unwanted size, we performed a double size selection  
807 using AMPure XP beads (Beckman Coulter) according to the manufacturer's  
808 instructions. The final fragmented and size-selected plasmid DNA was quality assessed  
809 with a DNA 1000 chip on an Agilent Bioanalyzer. Illumina library preparation involves a  
810 cascade of enzymatic reactions, each followed by a cleanup step with AMPure XP  
811 beads. Fragmentation generates plasmid DNA library with a mixture of blunt ends and  
812 5' and 3' overhangs. End repair, A-tailing, and adapter ligation reactions were  
813 performed on the fragmented DNA using the NEBNext DNA Library preparation kit for  
814 Illumina (New England Biolabs), according to the manufacturer's recommended  
815 protocols. For the adapter ligation, we used 0.5 ul of a 15uM double-stranded Y  
816 adapter, prepared by annealing Mod2\_TS\_Univ (ACGCTCTTCCGATC\*T) and  
817 Mod2\_Truseq (Phos-GATCGGAAGAGCACACGTCTGAACTCCAGTCA). In the  
818 preceding oligonucleotides, the asterisk and Phos represent phosphorothioate and 5'  
819 phosphate modifications, respectively.

820  
821 To specifically amplify UP barcodes and neighboring genomic fragment terminus by  
822 PCR, we used the UP-tag-specific primer oFAB2923\_Nspacer\_barseq\_universal, and  
823 P7\_MOD\_TS\_index1 primer. For the DOWN-tag amplification we used oFAB2924\_  
824 Nspacer\_barseq\_universal and P7\_MOD\_TS\_index2 primer. For the BAGseq UP  
825 barcode and DOWN barcode site enriching PCR, we used JumpStart Taq DNA  
826 polymerase (Sigma) in a 100 ul total volume with the following PCR program: 94°C for 2  
827 minutes and 25 cycles of 94°C 30 seconds, 65°C for 20 seconds, and 72°C for 30  
828 seconds, followed by a final extension at 72°C for 10 minutes. The final PCR product  
829 was purified using AMPure XP beads according to the manufacturer's instructions,  
830 eluted in 25 ul of water, and quantified on an Agilent Bioanalyzer with a DNA-1000 chip.  
831 Each BAGseq library was then sequenced on the HiSeq 2500 system (Illumina) with a  
832 150 SE run to map UP and DOWN barcodes to genomic inserts in the Dub-seq *E. coli*  
833 library.

### 834 835 **BAGseq data analysis**

836 BAGSeq reads were analyzed with *bagseq* script from the *Dub-seq* python library with  
837 default parameters (code available at <https://github.com/psnovichkov/DubSeq>). Fastq  
838 files for UP and DOWN barcodes with associated (cloned) genomic fragments are  
839 processed separately. For each read, the script looks for the flanking sequences around  
840 a barcode and requires an exact match of 9 nucleotides on both sides and a minimum  
841 quality score of 20 for each nucleotide in a barcode. The sequence downstream of the  
842 identified barcode is considered to be a candidate genomic fragment and is required to

843 be at least 15 nucleotides long for further processing. As a result, the initial list of the  
844 extracted barcodes and candidate genomic fragments is constructed.

845  
846 All extracted genomic fragments were compared to the *E. coli* genome sequence with  
847 BLAT using default parameters. Only hits with alignment block size of at least 15  
848 nucleotides and at most one indel were considered. It is also required that the extracted  
849 genomic fragment is mapped to one location in the genome. Thus, mappings to repeat  
850 regions were ignored. We applied two additional filters to minimize the number of  
851 erroneous associations between barcode and genomic location. First, we applied the  
852 same type of filter that we use for the analysis of BPSeq reads to filter out barcodes with  
853 a 1-nucleotide error.

854  
855 Second, the same barcode can be associated with different genomic fragments  
856 because of PCR artefacts (chimeras) or because multiple fragments were cloned  
857 between the same pair of barcodes. To filter out erroneous barcode mappings, the  
858 number of reads supporting different locations for the same barcode were calculated.  
859 To distinguish the true location from the false one, the frequency of the most abundant  
860 location (the number of supported reads) was compared with frequencies of all other  
861 locations for the same barcode. A given association between the barcode and the  
862 genomic location is considered to be true if the barcode does not have any other  
863 associated locations or the abundance of this association is at least two times more  
864 frequent than any other associations for the same barcode. As a result, the reference  
865 set of associations between UP (and separately for DOWN) barcodes and genomic  
866 locations is created, which we call 'BAGseq reference set'.

867  
868 The BPseq reference set of barcode pairs and BAGseq reference set are combined  
869 together to associate pairs of barcodes with genomic regions (to create the final 'Dub-  
870 seq reference set'). This step is done using the *bpag* script from the *Dub-seq* python  
871 library with default parameters. For each BPseq barcode pair, the script checks if the  
872 associations between UP and DOWN barcodes with genomic locations are present in  
873 the BAGSeq reference set. If both UP and DOWN barcodes (from BPseq reference set)  
874 are mapped to the genome, then the script checks the length of the region between the  
875 mapped locations and requires it to be between 100 nt and 6 kb. As a result, the final  
876 Dub-seq reference list of barcode pairs associated with genomic regions is created.  
877 Among total 10,600,088 reads for UP barcodes, usable reads were 3,884,931 (BAGseq  
878 UP barcode reads supporting the Dub-seq reference set), representing about 36.65% of  
879 total reads, whereas for total 9,671,635 reads for DOWN barcodes, usable reads were  
880 2,499,399, representing about 25.84% of total reads (BAGseq DOWN barcode reads  
881 supporting the Dub-seq reference set).

882

## 883 **Competitive growth experiments:**

884 For genome-wide competitive growth experiments, a single aliquot of the Dub-seq  
885 library in *E. coli* DH10B was thawed, inoculated into 25 ml of LB medium supplemented  
886 with chloramphenicol (30 ug/ml) and grown to mid-log phase. At mid-log phase, we  
887 collected cell pellets as a common reference for BarSeq (termed start or time-zero  
888 samples) and we used the remaining cells to set up competitive fitness assays under  
889 different experimental conditions at a starting OD600 of 0.02. For carbon source growth  
890 experiments, we used M9 defined medium supplemented with 0.3 mM L-leucine (as  
891 DH10B is auxotrophic for L-leucine)<sup>47</sup> and chloramphenicol. For experiments with stress  
892 compounds, we used an inhibitory but sublethal concentration of each compound, as  
893 determined previously<sup>15</sup>. All stress experiments were done in LB with chloramphenicol.  
894 All pooled fitness experiments were performed in 24-well microplates with 1.2 mL of  
895 media per well and grown in a multitron shaker. We took OD readings periodically in a  
896 Tecan M1000 instrument to ensure that the cells were growing and to confirm growth  
897 inhibition for the stress experiments. The assayed Dub-seq library cell pellets were  
898 stored at -80C prior to plasmid DNA extraction.

899

## 900 **BarSeq**

901 Plasmid DNA from Dub-seq library samples was extracted either individually using the  
902 Plasmid miniprep kit (Qiagen) or in 96-well format with a QIAprep 96 Turbo miniprep kit  
903 (Qiagen). Plasmid DNA was quantified with the Quant-iT dsDNA BR assay kit  
904 (Invitrogen). The BarSeq PCR of UP barcodes was done as previously described<sup>13</sup> with  
905 ~50 ng of plasmid template per BarSeq PCR reaction. To quantify the reproducibility of  
906 both UP and DOWN barcodes in competitive growth experiments, we collected plasmid  
907 DNA from nickel and cobalt experiments, and amplified both UP and DOWN barcodes  
908 in two separate PCRs using the same plasmid library template. For BarSeq PCR of  
909 DOWN barcodes, we used universal-forward-primer DT\_BarSeq\_p1\_FW and reverse  
910 primer DT\_BarSeq\_IT017. The PCR cycling conditions and purification steps were  
911 same as for the UP barcodes<sup>13</sup>. All experiments done on the same day and sequenced  
912 on the same lane are considered as a 'set'.

913

## 914 **BarSeq data analysis and fragment score calculation**

915 From HiSeq 4000 runs we obtained ~400 million of reads per lane, or 4.2 million reads  
916 per sample (for multiplexing 96 samples) typically >60% reads were informative after  
917 filtering out reads for sequencing errors and unmapped barcodes. BarSeq reads were  
918 analyzed with *barseq* script from the *Dub-seq* python library with default parameters.  
919 For each read, the script looks for the flanking sequences around each barcode and  
920 requires an exact match of 9 nucleotides on both sides and a minimum quality score of  
921 20 for each nucleotide in a barcode. The number of reads supporting each barcode is  
922 calculated. We apply the same type of filter that we use for the analysis of BPSeq reads

923 to filter out barcodes with single nucleotide substitutions relative to real barcodes (see  
924 BPSeq section). As a result, the list of barcode and their counts is created.

925

### 926 **Calculation of fragment scores (fScores)**

927 Given a reference list of barcodes mapped to the genomic regions (BPSeq and  
928 BAGSeq), and their counts in each sample (BarSeq), we estimate fitness values of each  
929 genomic fragment (strain) using *fscore* script from the Dub-seq python library with  
930 default parameters. First, the script identifies a subset of barcodes mapped to the  
931 genomic regions that are well represented in the time-zero samples for a given  
932 experiment set. We require that a barcode have at least 10 reads in at least one time-  
933 zero sample to be considered a valid barcode for a given experiment set. Then the  
934 *fscore* script calculates fitness score only for the strains with valid barcodes.

935

936 Strain fitness ( $f_i$ ) is calculated as a normalized  $\log_2$  ratio of counts between the  
937 treatment (condition or end) sample  $s_i$  and sum of counts across all (start) time-zero  $t_i$

938

$$939 \quad f_i = \log_2\left(\frac{s_{i+1}}{t_{i+1}}\right)$$

940

941 Then the strain fitness scores are normalized so that the median in each experiment is  
942 zero.

943

### 944 **Calculating gene-score (gScore)**

945 Given the fitness scores calculated for all Dub-seq fragments, we estimate a fitness  
946 score for each individual gene that is covered by at least one fragment. As mentioned in  
947 the Results, simply averaging the scores for the fragments that cover a gene gives  
948 spurious results for non-causative genes that are adjacent to a causative gene. To  
949 overcome this problem we modeled the fitness score of each fragment as the sum of  
950 the fitness scores of the genes that are completely covered by this fragment. Our model  
951 for estimating gene scores assumes that genes contribute independently to fitness, that  
952 most genes have little impact on fitness, and that intergenic regions have no effect on  
953 host fitness.

954

955 To estimate gene scores, we cannot use ordinary least squares (OLS), the most  
956 common type of regression, because of over fitting, which would produce unrealistic  
957 high positive and low negative scores for many genes. We also considered  
958 regularization methods (Ridge, LASSO, and ElasticNet), but these suffered from either  
959 too much shrinkage of fitness scores (biasing them towards zero) or failed to eliminate  
960 over fitting (see **Supplementary note**). Instead, we use Non-Negative Least Squares  
961 (NNLS) regression<sup>71</sup>, where the predicted gene scores are restricted to take only  
962 nonnegative values. If a gene with a potential benefit is next to (but not covered by) a

963 fragment with negative fitness, most regression methods would inflate the benefit of the  
964 gene and assign a negative score to the nearby gene. NNLS instead ignores the (often  
965 noisy) negative scores for the nearby fragments. To estimate negative gene scores, we  
966 also used NNLS, but with the signs of the fragment scores flipped.

967

968 In our model, the expected fitness of a fragment is given by

$$f_i = \sum_j g_{ij}$$

969 where  $g_{ij}$  is a fitness score of a gene covered by  $i$ -th fragment completely. The NNLS  
970 minimizes

971

$$\|Ag - f\|_2^2, \text{ subject to } g \geq 0$$

972

973 where  $g$  a vector of gene fitness scores to be estimated,  $f$  is vector of the “observed”  
974 fitness scores of fragments,  $A$  a matrix of ones and zeros defining which gene is  
975 covered by which fragment completely. Gene scores were calculated using the *gscore*  
976 script from the Dub-seq python library with default parameters, which uses the *nnls*  
977 function from the *optimize* package of the *scipy* python library.

## 978 **High-confidence gene scores and estimating the false discovery rate**

979 We used several filters to identify gene scores that were likely to be of high-confidence  
980 and reliable. Whereas the non-negative regression was used to determine if the high  
981 fitness of the fragments covering the gene are due to this gene or a nearby gene, these  
982 filters were intended to ensure that the fragments covering the gene had a genuine  
983 benefit. The first filter was  $|\text{gene score}| \geq 2$ , as such a large effect occurred just 4 times  
984 in 17 control comparisons between independently-processed but identical “start”  
985 samples (0.2 per experiment). In contrast the actual conditions gave 40 large effects per  
986 experiment on average (over 150 times more).

987

988 Second, we noticed that some genes had high scores because of a single fragment with  
989 a very high score. These fragments did not have high scores in replicate experiments,  
990 so their high scores might be due to secondary mutations. To filter out these cases, we  
991 performed a single-sample  $t$  test on the fragment scores (for the fragments that covered  
992 the gene) and required  $P < 0.05$ . This test asks if the mean is significantly different from  
993 a reference value. To handle uncertainty in the true centering of the fragment scores  
994 (which were normalized to have a median of zero), we considered the mean of all  
995 fragment scores for the experiment. We used this as the reference value (instead of  
996 zero) if this mean had the same sign as the gene’s score. This makes the filter slightly

997 more stringent. If the gene has just one fragment, then we cannot apply the  $t$  test, so we  
998 instead require that |fragment score| be in the top 1% for this experiment.

999  
1000 Third, we checked that the effect was larger relative to the expected noise in the mean  
1001 of the fragment scores that cover the gene. The expected noise for each fragment can  
1002 be estimated as  $\sqrt{1/(1+\text{count\_after}) + 1/(1+\text{count\_start})} / \ln(2)$ . This approximation is  
1003 derived from the best case that the noise in the counts follows a Poisson distribution.  
1004 The expected noise for the mean of the fragment scores is then  
1005  $\sqrt{\text{sum}(\text{fragment\_noise}^2)} / \text{nfragments}$ . Note that  $z = \text{mean}(\text{fragment score}) / \text{noise}$   
1006 would (ideally) follow the standard normal distribution. We use  $|z| \geq 4$  as a filter; with  
1007 4,303 genes being assayed, we would expect about 0.3 false positives per experiment.

1008  
1009 "Filtered effects" (that passed all three filters) were considered to be reliable. Reliable  
1010 effects were considered to be high-confidence if the gene was covered by multiple  
1011 fragments. Because of the risk of secondary mutations, a measurement for a gene with  
1012 a single fragment was only considered high-confidence if it was reliable and was also  
1013 supported by a large effect ( $|\text{score}| \geq 2$ ) in another experiment for that compound.

1014  
1015 The filtered effects were usually consistent across replicate experiments and represent  
1016 reliable scores. We had two biological replicates for 64 of the 82 conditions (a  
1017 compound at a given concentration) that we studied. Across these 64 pairs of replicate  
1018 experiments, 85% of genes with filtered effects in one replicate were consistent ( $|\text{score}|$   
1019  $\geq 1.5$  and the same sign) in the other replicate. Large effects ( $|\text{score}| \geq 2$ ) were more  
1020 likely to replicate if they were filtered (85% vs. 59% otherwise). Among filtered effects  
1021 for genes covered by more than one fragment, 39% of the effects that did not replicate  
1022 were from a single condition (zinc sulfate stress at 1 mM). We did not identify any  
1023 obvious issue for the data from this condition. In total, 4,303 genes are covered by at  
1024 least one fragment, but there are only 4,151 genes with at least one gene score  
1025 (adequate representation in at least one start sample).

1026  
1027 To estimate the false discovery rate for high-confidence effects, we randomly shuffled  
1028 the mapping of barcodes to fragments, recomputed the mean scores for each gene in  
1029 each experiment, and identified high-confidence effects as for the genuine data. This  
1030 shuffling test will probably overestimate the FDR because it assumes that all of the  
1031 variability in the fragment scores is due to noise. Also, we used the mean score, rather  
1032 than regression-based gene score, in this test. This might also lead to an overestimate  
1033 of the FDR. We repeated the shuffle procedure 10 times. On average, each shuffled  
1034 data set had 75 high-confidence effects, while the actual data had 4,051 high-  
1035 confidence effects, so we estimated the false discovery rate as  $75/4051 = 1.9\%$ .

1036

### 1037 **Calculating gene-pair fitness score**

1038 Although our model assumes that the genes on a fragment contribute independently to  
1039 fitness, there are cases where multiple nearby genes work together to confer a  
1040 phenotype. For estimating such ‘epistatic’ synergistic fitness contribution by neighboring  
1041 pair of genes, we included additional variables in our fitness calculation to account for  
1042 the contribution of pairs of adjacent genes (and their intergenic regions). For a gene-pair  
1043 to qualify to be valid hit, the score for the gene-pair has to be more than the individual  
1044 gene scores from single-gene regression model, scores should be consistent across  
1045 replicates and should be supported by more than one fragment. After manual filtering,  
1046 we found 6 high scoring epistatic-effect instances where gene-pairs positively contribute  
1047 to the host fitness under specific condition (**Supplementary Table 5**). Among these, 3  
1048 gene-pairs have related functions (*fetA-fetB* on nickel, *ampD-ampE* on benzethonium,  
1049 *ackA-pta* on D-lactate<sup>48</sup>) and make biological sense. However, in the other 3 high  
1050 scoring gene-pairs *arcA-yjjY*, *hns-tdk* and *yfiF-trxC*, each gene is divergently transcribed  
1051 and the reason behind combined fitness phenotype is not obvious. We speculate, the  
1052 fitness phenotype in these cases may be function of intergenic regions in addition to the  
1053 encoded genes.

1054

### 1055 **Experimental validation of single genes**

1056 To experimentally validate some of top hits in our Dub-seq results we used the ASKA  
1057 ORF collection<sup>29</sup>. The ASKA library consists of *E. coli* ORFs cloned on a pMB1  
1058 replication origin plasmid and driven by an IPTG-inducible promoter. We extracted  
1059 individual ASKA ORF plasmids from the collection, sequence confirmed and  
1060 transformed the plasmids into our assay strain *E. coli* DH10B. As the plasmid copy  
1061 number and the strength of promoter and ribosome binding site used in the ASKA ORF  
1062 collection is different from the broad-host pBBR1 plasmid system used in *E. coli* Dub-seq  
1063 library, we screened for an optimum IPTG levels to induce the expression of specific  
1064 gene in order to study the host fitness. We grew the individual strains in 96-well  
1065 microplates with 150  $\mu$ L total volume per well. These plates were grown at 30°C with  
1066 shaking in a Tecan microplate reader (either Sunrise or Infinite F200) with optical  
1067 density readings every 15 minutes.

1068

### 1069 **Library visualization tools**

1070 We used the Dub-seq viewer tool from the *Dub-seq* python library  
1071 (<https://github.com/psnovichkov/DubSeq>) to generate regions of the *E. coli* chromosome  
1072 covering fragments (landscape mode) presented in **Fig 2a**. To generate fitness score  
1073 plots as shown in **Fig. 3a and 3b**, and **Supplement Figs. 4, 6 and 7**, we used gene-  
1074 browser mode. We used Circa software (OmGenomics) to generate genome coverage  
1075 plot shown in **Fig. 2a**.

1076

1077 **Code and metadata availability**

1078 Code for processing and analyzing Dub-seq data is available at

1079 <https://github.com/psnovichkov/DubSeq>

1080

1081 Complete data from all experiments (read counts per barcode, fragment scores and

1082 gene scores) is deposited here: <https://doi.org/10.6084/m9.figshare.6752753.v1>

1083

1084 Link to website with supplementary information and bulk data downloads:

1085 <http://morgannprice.org/dubseq18/>

1086

1087

1088 **REFERENCES**

1089

1090 1 Markowitz, V. M. *et al.* Ten years of maintaining and expanding a microbial genome  
1091 and metagenome analysis system. *Trends Microbiol* **23**, 730-741,  
1092 doi:10.1016/j.tim.2015.07.012 (2015).

1093 2 Chang, Y. C. *et al.* COMBREX-DB: an experiment centered database of protein  
1094 function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res* **44**, D330-  
1095 335, doi:10.1093/nar/gkv1324 (2016).

1096 3 Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public  
1097 databases: misannotation of molecular function in enzyme superfamilies. *PLoS*  
1098 *Comput Biol* **5**, e1000605, doi:10.1371/journal.pcbi.1000605 (2009).

1099 4 Blaser, M. J. *et al.* Toward a Predictive Understanding of Earth's Microbiomes to  
1100 Address 21st Century Challenges. *MBio* **7**, doi:10.1128/mBio.00714-16 (2016).

1101 5 Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout  
1102 mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008, doi:10.1038/msb4100050  
1103 (2006).

1104 6 Koo, B. M. *et al.* Construction and Analysis of Two Genome-Scale Deletion Libraries  
1105 for Bacillus subtilis. *Cell Syst* **4**, 291-305 e297, doi:10.1016/j.cels.2016.12.013  
1106 (2017).

1107 7 Giaever, G. & Nislow, C. The yeast deletion collection: a decade of functional  
1108 genomics. *Genetics* **197**, 451-465, doi:10.1534/genetics.114.161620 (2014).

1109 8 Barker, C. A., Farha, M. A. & Brown, E. D. Chemical genomic approaches to study  
1110 model microbes. *Chem Biol* **17**, 624-632, doi:10.1016/j.chembiol.2010.05.010  
1111 (2010).

1112 9 Brochado, A. R. & Typas, A. High-throughput approaches to understanding gene  
1113 function and mapping network architecture in bacteria. *Curr Opin Microbiol* **16**, 199-  
1114 206, doi:10.1016/j.mib.2013.01.008 (2013).

1115 10 Wang, H. H. *et al.* Programming cells by multiplex genome engineering and  
1116 accelerated evolution. *Nature* **460**, 894-898, doi:10.1038/nature08187 (2009).

1117 11 Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. & Gill, R. T. Rapid  
1118 profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat*  
1119 *Biotechnol* **28**, 856-862, doi:10.1038/nbt.1653 (2010).



- 1120 12 van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing  
1121 for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6**, 767-  
1122 772, doi:10.1038/nmeth.1377 (2009).
- 1123 13 Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by  
1124 sequencing randomly bar-coded transposons. *MBio* **6**, e00306-00315,  
1125 doi:10.1128/mBio.00306-15 (2015).
- 1126 14 Peters, J. M. *et al.* A Comprehensive, CRISPR-based Functional Analysis of Essential  
1127 Genes in Bacteria. *Cell* **165**, 1493-1506, doi:10.1016/j.cell.2016.05.003 (2016).
- 1128 15 Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown  
1129 function. *Nature* **557**, 503-509, doi:10.1038/s41586-018-0124-0 (2018).
- 1130 16 Prelich, G. Gene overexpression: uses, mechanisms, and interpretation. *Genetics* **190**,  
1131 841-854, doi:10.1534/genetics.111.136911 (2012).
- 1132 17 Sandegren, L. & Andersson, D. I. Bacterial gene amplification: implications for the  
1133 evolution of antibiotic resistance. *Nat Rev Microbiol* **7**, 578-588,  
1134 doi:10.1038/nrmicro2174 (2009).
- 1135 18 Elliott, K. T., Cuff, L. E. & Neidle, E. L. Copy number change: evolving views on gene  
1136 amplification. *Future Microbiol* **8**, 887-899, doi:10.2217/fmb.13.53 (2013).
- 1137 19 Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. Targeted selection of recombinant  
1138 clones through gene dosage effects. *Proc Natl Acad Sci U S A* **80**, 6750-6754 (1983).
- 1139 20 Ho, C. H. *et al.* A molecular barcoded yeast ORF library enables mode-of-action  
1140 analysis of bioactive compounds. *Nat Biotechnol* **27**, 369-377, doi:10.1038/nbt.1534  
1141 (2009).
- 1142 21 Soo, V. W., Hanson-Manful, P. & Patrick, W. M. Artificial gene amplification reveals an  
1143 abundance of promiscuous resistance determinants in *Escherichia coli*. *Proc Natl*  
1144 *Acad Sci U S A* **108**, 1484-1489, doi:10.1073/pnas.1012108108 (2011).
- 1145 22 Hoegler, K. J. & Hecht, M. H. Artificial Gene Amplification in *Escherichia coli* Reveals  
1146 Numerous Determinants for Resistance to Metal Toxicity. *J Mol Evol* **86**, 103-110,  
1147 doi:10.1007/s00239-018-9830-3 (2018).
- 1148 23 Qimron, U., Marintcheva, B., Tabor, S. & Richardson, C. C. Genomewide screens for  
1149 *Escherichia coli* genes affecting growth of T7 bacteriophage. *Proc Natl Acad Sci U S A*  
1150 **103**, 19039-19044, doi:10.1073/pnas.0609428103 (2006).
- 1151 24 Li, X. *et al.* Multicopy suppressors for novel antibacterial compounds reveal targets  
1152 and drug efflux susceptibility. *Chem Biol* **11**, 1423-1430,  
1153 doi:10.1016/j.chembiol.2004.08.014 (2004).
- 1154 25 Patrick, W. M., Quandt, E. M., Swartzlander, D. B. & Matsumura, I. Multicopy  
1155 suppression underpins metabolic evolvability. *Mol Biol Evol* **24**, 2716-2722,  
1156 doi:10.1093/molbev/msm204 (2007).
- 1157 26 Lynch, M. D., Warnecke, T. & Gill, R. T. SCALES: multiscale analysis of library  
1158 enrichment. *Nat Methods* **4**, 87-93, doi:10.1038/nmeth946 (2007).
- 1159 27 Nicolaou, S. A., Gaida, S. M. & Papoutsakis, E. T. Coexisting/Coexpressing Genomic  
1160 Libraries (CoGeL) identify interactions among distantly located genetic loci for  
1161 developing complex microbial phenotypes. *Nucleic Acids Res* **39**, e152,  
1162 doi:10.1093/nar/gkr817 (2011).
- 1163 28 Dunlop, M. J. *et al.* Engineering microbial biofuel tolerance and export using efflux  
1164 pumps. *Mol Syst Biol* **7**, 487, doi:10.1038/msb.2011.21 (2011).

- 1165 29 Kitagawa, M. *et al.* Complete set of ORF clones of Escherichia coli ASKA library (a  
1166 complete set of E. coli K-12 ORF archive): unique resources for biological research.  
1167 *DNA Res* **12**, 291-299, doi:10.1093/dnares/dsi012 (2005).
- 1168 30 Wang, H. H. *et al.* Genome-scale promoter engineering by coselection MAGE. *Nat*  
1169 *Methods* **9**, 591-593, doi:10.1038/nmeth.1971 (2012).
- 1170 31 Freed, E. F. *et al.* Genome-Wide Tuning of Protein Expression Levels to Rapidly  
1171 Engineer Microbial Traits. *ACS Synth Biol* **4**, 1244-1253,  
1172 doi:10.1021/acssynbio.5b00133 (2015).
- 1173 32 Judson, N. & Mekalanos, J. J. TnAraOut, a transposon-based approach to identify and  
1174 characterize essential bacterial genes. *Nat Biotechnol* **18**, 740-745,  
1175 doi:10.1038/77305 (2000).
- 1176 33 Dong, C., Fontana, J., Patel, A., Carothers, J. M. & Zalatan, J. G. Synthetic CRISPR-Cas  
1177 gene activators for transcriptional reprogramming in bacteria. *Nat Commun* **9**, 2489,  
1178 doi:10.1038/s41467-018-04901-6 (2018).
- 1179 34 Leis, B., Angelov, A. & Liebl, W. Screening and expression of genes from  
1180 metagenomes. *Adv Appl Microbiol* **83**, 1-68, doi:10.1016/B978-0-12-407678-  
1181 5.00001-5 (2013).
- 1182 35 Ekkers, D. M., Cretoiu, M. S., Kielak, A. M. & Elsas, J. D. The great screen anomaly--a  
1183 new frontier in product discovery through functional metagenomics. *Appl Microbiol*  
1184 *Biotechnol* **93**, 1005-1020, doi:10.1007/s00253-011-3804-3 (2012).
- 1185 36 Uchiyama, T. & Miyazaki, K. Functional metagenomics for enzyme discovery:  
1186 challenges to efficient screening. *Curr Opin Biotechnol* **20**, 616-622,  
1187 doi:10.1016/j.copbio.2009.09.010 (2009).
- 1188 37 Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the  
1189 antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128-1131,  
1190 doi:10.1126/science.1176950 (2009).
- 1191 38 Munck, C. *et al.* Limited dissemination of the wastewater treatment plant core  
1192 resistome. *Nat Commun* **6**, 8452, doi:10.1038/ncomms9452 (2015).
- 1193 39 Yaung, S. J. *et al.* Improving microbial fitness in the mammalian gut by in vivo  
1194 temporal functional metagenomics. *Molecular Systems Biology* **11**, 788-788,  
1195 doi:10.15252/msb.20145866 (2015).
- 1196 40 Gibson, M. K. *et al.* Developmental dynamics of the preterm infant gut microbiota  
1197 and antibiotic resistome. *Nat Microbiol* **1**, 16024, doi:10.1038/nmicrobiol.2016.24  
1198 (2016).
- 1199 41 Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome*  
1200 *Res* **19**, 1836-1842, doi:10.1101/gr.093955.109 (2009).
- 1201 42 Studier, F. W. & Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct  
1202 selective high-level expression of cloned genes. *J Mol Biol* **189**, 113-130 (1986).
- 1203 43 Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal  
1204 gene transfer. *Science* **318**, 1449-1452, doi:10.1126/science.1147112 (2007).
- 1205 44 Oh, J. *et al.* A universal TagModule collection for parallel genetic analysis of  
1206 microorganisms. *Nucleic Acids Res* **38**, e146, doi:10.1093/nar/gkq419 (2010).
- 1207 45 Rodrigue, A., Effantin, G. & Mandrand-Berthelot, M. A. Identification of rcnA (yohM),  
1208 a nickel and cobalt resistance gene in Escherichia coli. *J Bacteriol* **187**, 2912-2916,  
1209 doi:10.1128/JB.187.8.2912-2916.2005 (2005).

- 1210 46 Romero, D. & Palacios, R. Gene amplification and genomic plasticity in prokaryotes.  
1211 *Annu Rev Genet* **31**, 91-111, doi:10.1146/annurev.genet.31.1.91 (1997).
- 1212 47 Durfee, T. *et al.* The complete genome sequence of Escherichia coli DH10B: insights  
1213 into the biology of a laboratory workhorse. *J Bacteriol* **190**, 2597-2606,  
1214 doi:10.1128/JB.01695-07 (2008).
- 1215 48 Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about  
1216 Escherichia coli K-12. *Nucleic Acids Res* **45**, D543-D550, doi:10.1093/nar/gkw1003  
1217 (2017).
- 1218 49 Egler, M., Grosse, C., Grass, G. & Nies, D. H. Role of the extracytoplasmic function  
1219 protein family sigma factor RpoE in metal resistance of Escherichia coli. *J Bacteriol*  
1220 **187**, 2297-2307, doi:10.1128/JB.187.7.2297-2307.2005 (2005).
- 1221 50 Grabowicz, M. & Silhavy, T. J. Envelope Stress Responses: An Interconnected Safety  
1222 Net. *Trends Biochem Sci* **42**, 232-242, doi:10.1016/j.tibs.2016.10.002 (2017).
- 1223 51 Nishino, K., Yamasaki, S., Hayashi-Nishino, M. & Yamaguchi, A. Effect of NlpE  
1224 overproduction on multidrug resistance in Escherichia coli. *Antimicrob Agents*  
1225 *Chemother* **54**, 2239-2243, doi:10.1128/AAC.01677-09 (2010).
- 1226 52 Guo, M. S. *et al.* MicL, a new sigmaE-dependent sRNA, combats envelope stress by  
1227 repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes Dev* **28**,  
1228 1620-1634, doi:10.1101/gad.243485.114 (2014).
- 1229 53 Freeman, J. L., Persans, M. W., Nieman, K. & Salt, D. E. Nickel and cobalt resistance  
1230 engineered in Escherichia coli by overexpression of serine acetyltransferase from  
1231 the nickel hyperaccumulator plant *Thlaspi goesingense*. *Appl Environ Microbiol* **71**,  
1232 8627-8633, doi:10.1128/AEM.71.12.8627-8633.2005 (2005).
- 1233 54 Lim, B. *et al.* RNase III controls the degradation of corA mRNA in Escherichia coli. *J*  
1234 *Bacteriol* **194**, 2214-2220, doi:10.1128/JB.00099-12 (2012).
- 1235 55 Couce, A. *et al.* Genomewide overexpression screen for fosfomycin resistance in  
1236 Escherichia coli: MurA confers clinical resistance at low fitness cost. *Antimicrob*  
1237 *Agents Chemother* **56**, 2767-2769, doi:10.1128/AAC.06122-11 (2012).
- 1238 56 Li, H., Zhang, D. F., Lin, X. M. & Peng, X. X. Outer membrane proteomics of kanamycin-  
1239 resistant Escherichia coli identified MipA as a novel antibiotic resistance-related  
1240 protein. *FEMS Microbiol Lett* **362**, doi:10.1093/femsle/fnv074 (2015).
- 1241 57 Silver, L. L. Fosfomycin: Mechanism and Resistance. *Cold Spring Harb Perspect Med*  
1242 **7**, doi:10.1101/cshperspect.a025262 (2017).
- 1243 58 Nicolaou, S. A., Fast, A. G., Nakamaru-Ogiso, E. & Papoutsakis, E. T. Overexpression of  
1244 fetA (ybbL) and fetB (ybbM), Encoding an Iron Exporter, Enhances Resistance to  
1245 Oxidative Stress in Escherichia coli. *Appl Environ Microbiol* **79**, 7210-7219,  
1246 doi:10.1128/AEM.02322-13 (2013).
- 1247 59 Hoon, S. *et al.* An integrated platform of genomic assays reveals small-molecule  
1248 bioactivities. *Nat Chem Biol* **4**, 498-506, doi:10.1038/nchembio.100 (2008).
- 1249 60 Thompson, K. M., Rhodius, V. A. & Gottesman, S. SigmaE regulates and is regulated  
1250 by a small RNA in Escherichia coli. *J Bacteriol* **189**, 4243-4256,  
1251 doi:10.1128/JB.00020-07 (2007).
- 1252 61 Shuman, H. A. & Silhavy, T. J. The art and design of genetic screens: Escherichia coli.  
1253 *Nat Rev Genet* **4**, 419-431, doi:10.1038/nrg1087 (2003).

- 1254 62 Grothe, S., Krogsrud, R. L., McClellan, D. J., Milner, J. L. & Wood, J. M. Proline transport  
1255 and osmotic stress response in *Escherichia coli* K-12. *J Bacteriol* **166**, 253-259  
1256 (1986).
- 1257 63 Paradis-Bleau, C., Kritikos, G., Orlova, K., Typas, A. & Bernhardt, T. G. A genome-wide  
1258 screen for bacterial envelope biogenesis mutants identifies a novel factor involved  
1259 in cell wall precursor metabolism. *PLoS Genet* **10**, e1004056,  
1260 doi:10.1371/journal.pgen.1004056 (2014).
- 1261 64 Neufeld, J. D. *et al.* Open resource metagenomics: a model for sharing metagenomic  
1262 libraries. *Stand Genomic Sci* **5**, 203-210, doi:10.4056/sigs.1974654 (2011).
- 1263 65 Pal, C. *et al.* Metal Resistance and Its Association With Antibiotic Resistance. *Adv*  
1264 *Microb Physiol* **70**, 261-313, doi:10.1016/bs.ampbs.2017.02.001 (2017).
- 1265 66 Gaida, S. M. *et al.* Expression of heterologous sigma factors enables functional  
1266 screening of metagenomic and heterologous genomic libraries. *Nat Commun* **6**,  
1267 7045, doi:10.1038/ncomms8045 (2015).
- 1268 67 Ausubel, F. M. *Short protocols in molecular biology : a compendium of methods from*  
1269 *Current protocols in molecular biology*. 5th edn, (Wiley, 2002).
- 1270 68 Lee, T. S. *et al.* BglBrick vectors and datasheets: A synthetic biology platform for  
1271 gene expression. *J Biol Eng* **5**, 12, doi:10.1186/1754-1611-5-12 (2011).
- 1272 69 Kovach, M. E., Phillips, R. W., Elzer, P. H., Roop, R. M., 2nd & Peterson, K. M.  
1273 pBBR1MCS: a broad-host-range cloning vector. *Biotechniques* **16**, 800-802 (1994).
- 1274 70 Sambrook, J., Russell, D. W. & Sambrook, J. *The condensed protocols from Molecular*  
1275 *cloning : a laboratory manual*. (Cold Spring Harbor Laboratory Press, 2006).
- 1276 71 Lawson, C. L., Hanson, R. J. & Society for Industrial and Applied Mathematics. in  
1277 *Classics in applied mathematics 15* 1 electronic text (xii, 337 p (Society for  
1278 Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6,  
1279 Philadelphia, PA 19104), Philadelphia, Pa., 1995).
- 1280
- 1281

1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304

## SUPPLEMENTARY INFORMATION

### Supplementary Tables:

**Supplementary Table 1.** List of 135 genes not represented in *E. coli* Dub-seq library

**Supplementary Table 2.** List of protein-coding genes with details on number of Dub-seq fragments covering the gene, and if the gene is essential (according to the Keio library<sup>5</sup>), has RB-TnSeq data<sup>15</sup> and has Dub-seq data (this work).

**Supplementary Table 3.** Filtered gene scores for reliable effects in Dub-seq dataset and if they have representative data in RB-TnSeq mutant library<sup>15</sup>

**Supplementary Table 4.** List of genes whose high dosage is known to yield positive fitness effects

**Supplementary Table 5.** Novel gene-function associations with fitness score  $\geq 4$ ; hypothesis and general notes

**Supplementary Table 6.** List of primers used in this work

**Supplementary Table 7.** List of plasmids used in this work

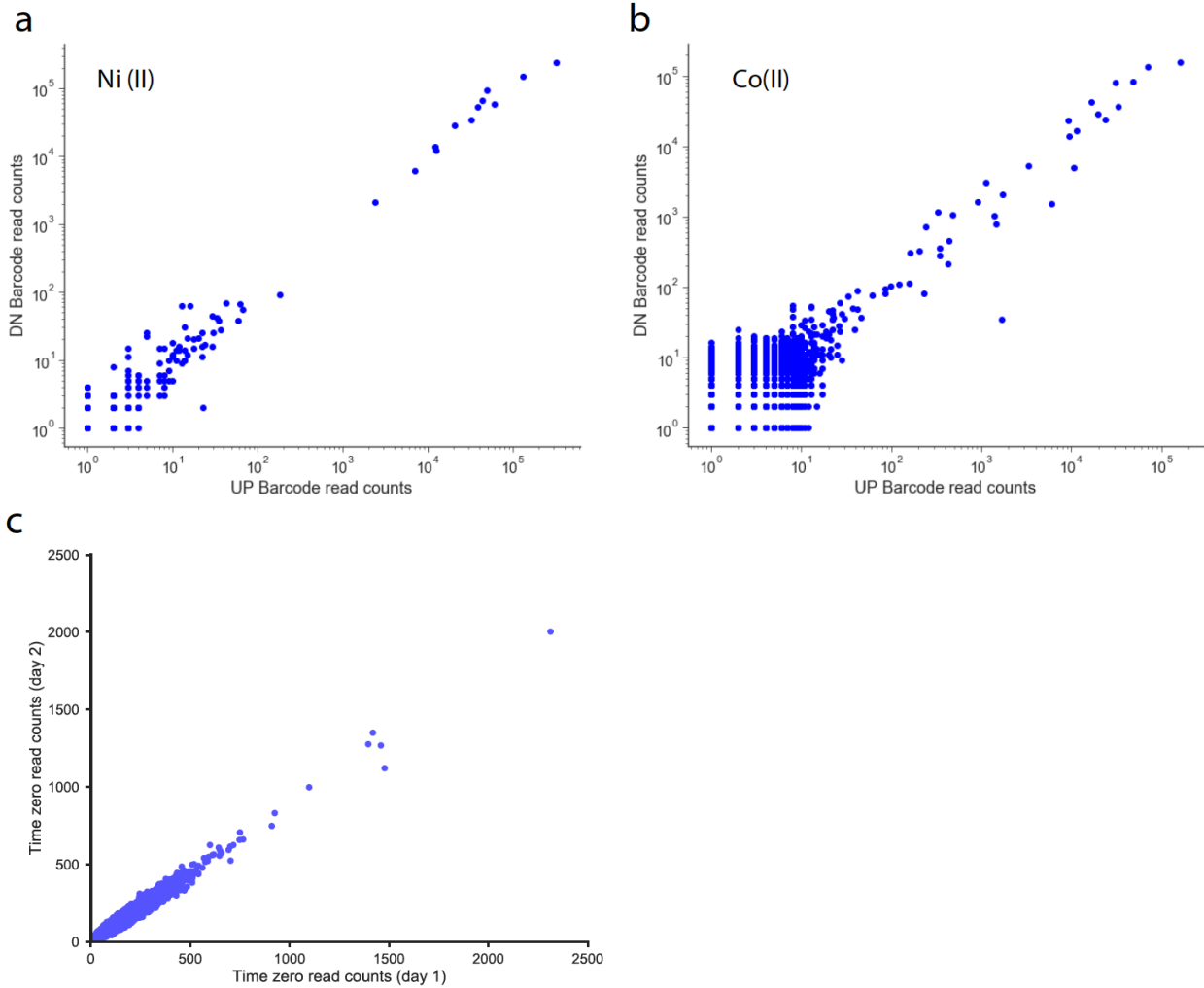
**Supplementary Table 8.** List of strains used in this work

Link to website with supplementary information:

<http://morgannprice.org/dubseq18/>

1305  
1306  
1307  
1308  
1309

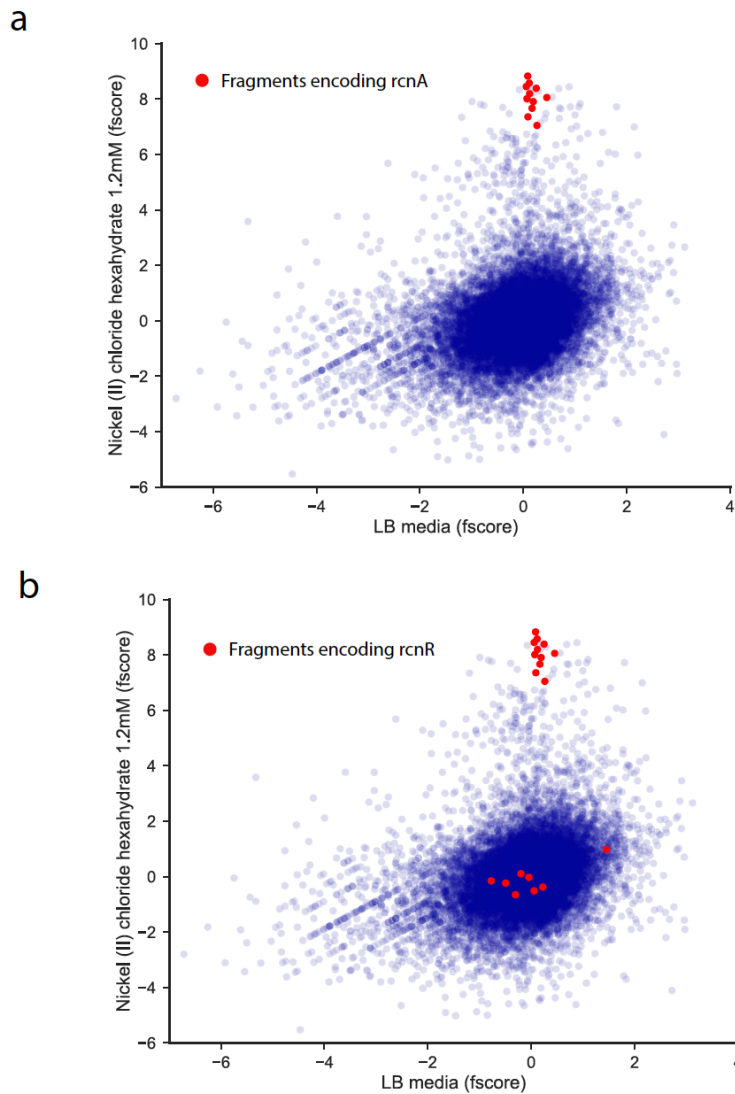
### Supplementary Figures:



1310  
1311  
1312  
1313  
1314  
1315  
1316

**Supplementary Fig. 1. BarSeq reproducibility:** Comparison of UP and DOWN barcode BarSeq reads for (a) Nickel and (b) Cobalt condition. (c) Comparison of UP barcode reads for two independent start (time-zero) samples.

1317



1318

1319

1320 **Supplementary Fig. 2. Fragment score comparisons:** Fragment score (fscore)

1321 comparisons for all fragments in LB (x-axis) and LB with nickel (y-axis). (a) Fragments

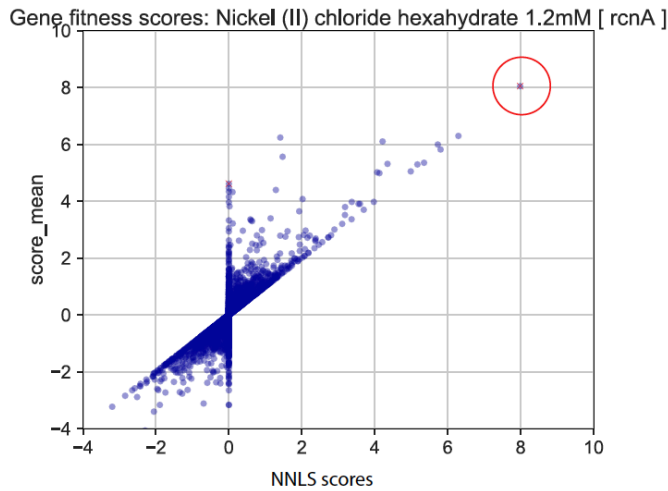
1322 fully covering *rcnA* are highlighted in red. (b) Fragments fully covering *rcnR* are

1323 highlighted in red.

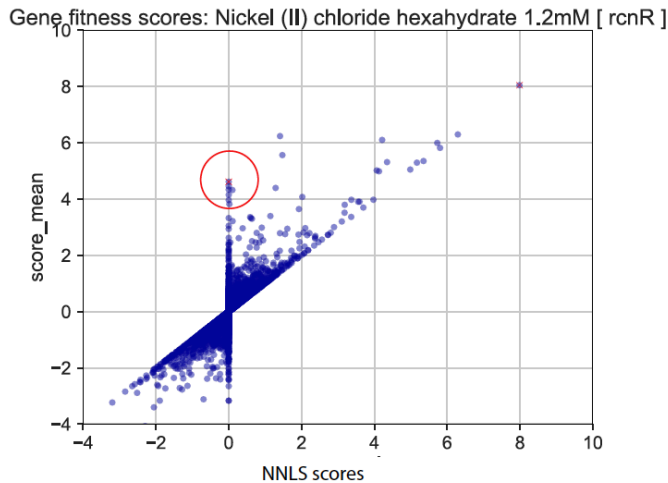
1324

1325  
1326  
1327

a



b

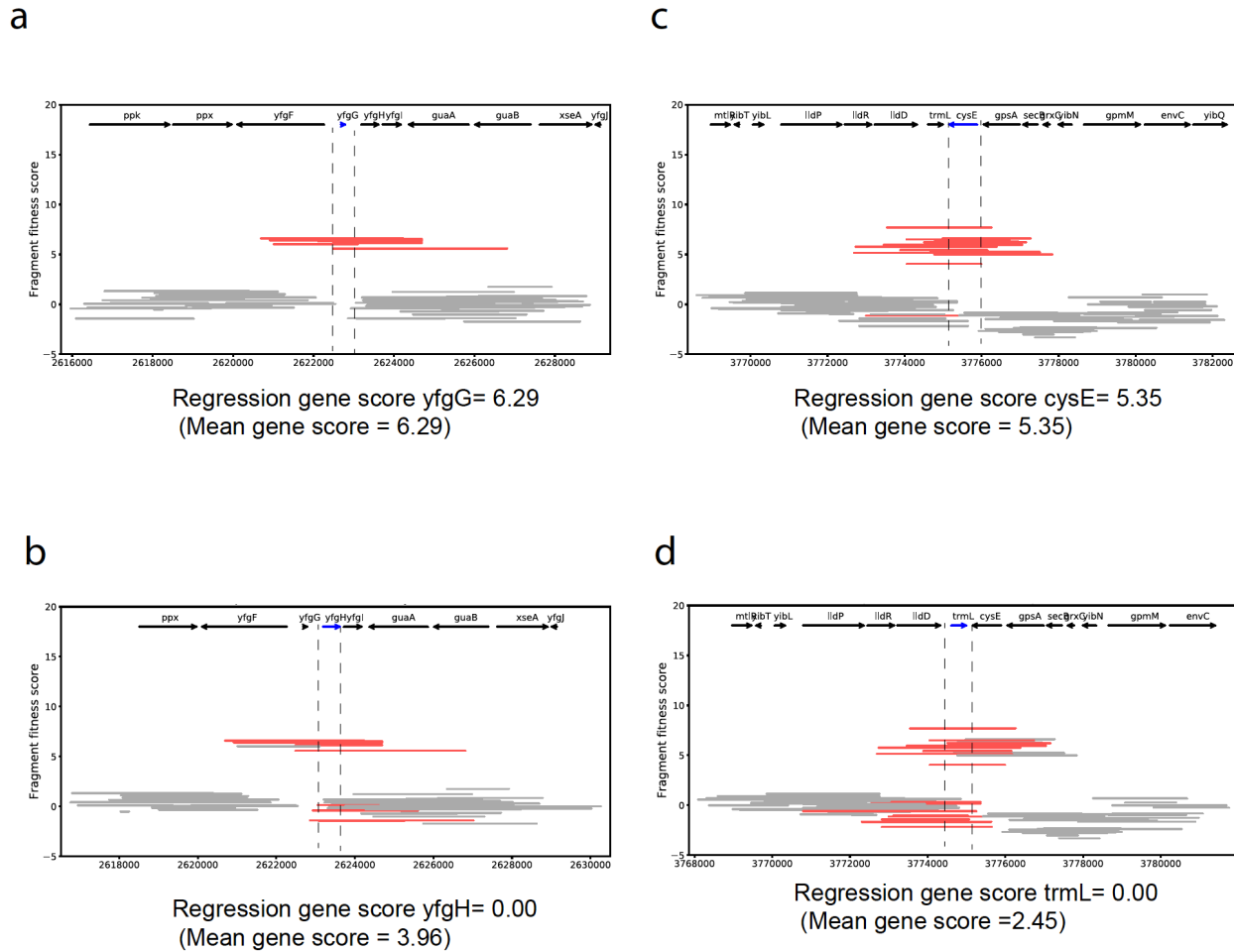


1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335

**Supplementary Fig. 3. Comparison of gene scores from regression analysis and mean gene scores:** Comparison between gene fitness scores calculated using Non-Negative Least Squares regression (NNLS) method and the mean score method under nickel stress (a) Fitness score for *rcnA* (red circle) (b) Fitness score for *rcnR* (red circle).



1336  
1337



1338

1339

1340

1341

1342

1343

1344

1345

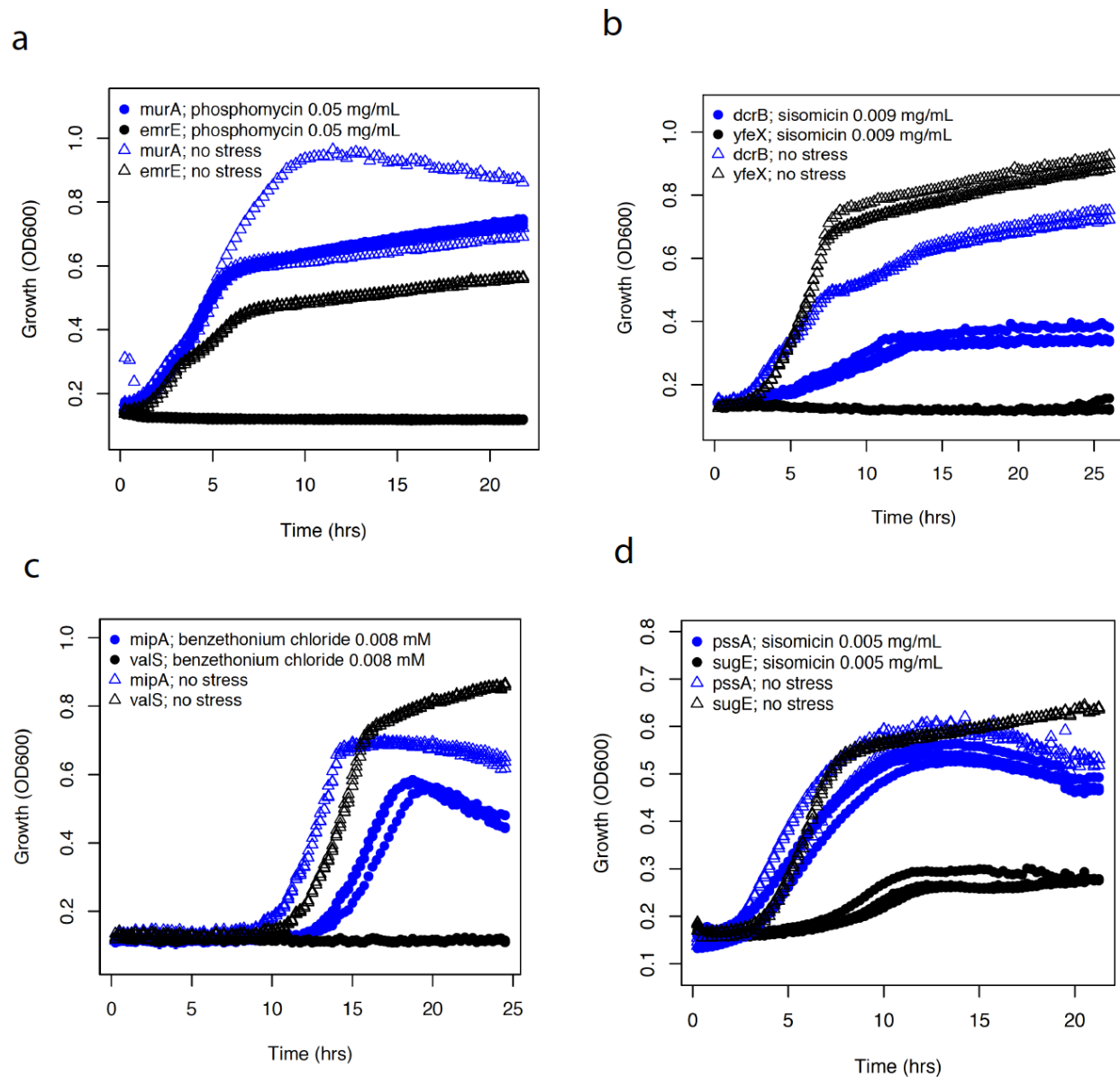
1346

1347

1348

**Supplementary Fig. 4. Fragment and gene Dub-seq scores:** Dub-seq fragment (strain) data for different regions under elevated nickel stress (y-axis). Each line shows a Dub-seq fragment with those that completely cover the indicated gene are in red. The mean and regression scores for each indicated gene are shown below each plot. Compare scores for (a) *yfgG* with (b) *yfgH*, and (c) *cysE* with (d) *trmL*. Note that the mean and regression scores for *yfgH* and *trmL* are different. The mean score is incorrectly high for *yfgH* and *trmL* and is due to the presence of *yfgG* and *cysE* on a number of fragments.

1349  
1350

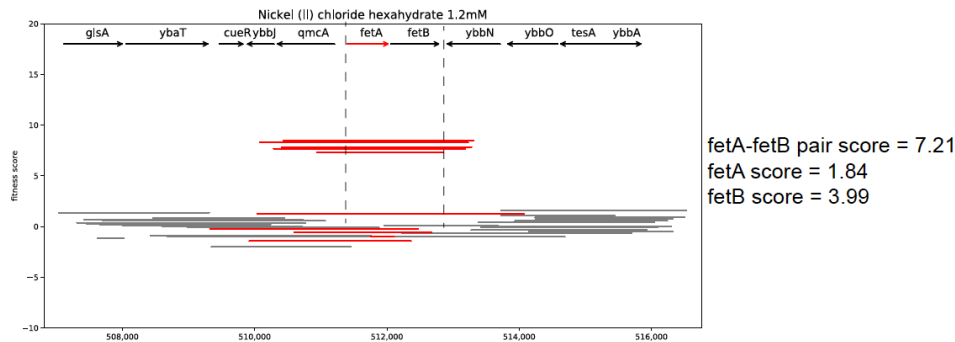


1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359

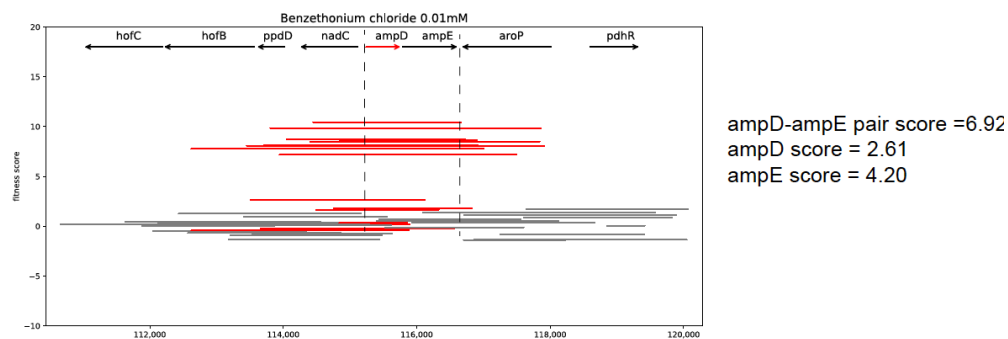
**Supplementary Fig. 5. Additional validation growth curves for Dub-seq high scoring genes.** (a) Growth of *E. coli* overexpressing *murA* under phosphomycin stress; *emrE* is a control. (b) Growth of *E. coli* overexpressing *dcrB* under sisomicin stress; *yfeX* is a control. (c) Growth of *E. coli* overexpressing *mipA* under benzethonium chloride stress; *valS* is used as a control. (d) Growth of *E. coli* overexpressing *pssA* under sisomicin stress; *sugE* is used as a control.

1360  
1361  
1362

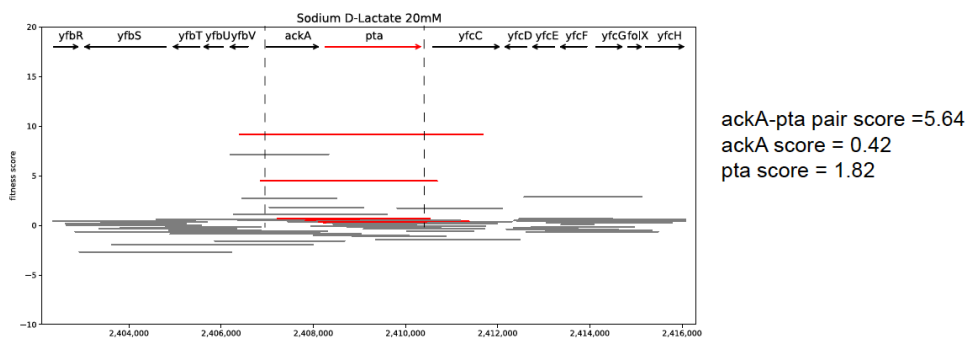
a



b



c



1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374

**Supplementary Fig. 6. Dub-seq gene-pair fitness scores:** Dub-seq fragment (strain) data (y-axis) for region surrounding gene-pair of interest (x-axis). The covered fragments are shown in red and partially covered gene-pair-neighborhood fragments are shown in gray. The regression scores each gene-pair of interest are shown next to each plot. Compare scores for (a) *fetA* and *fetB* with *fetA-fetB* pair with (b) *ampD* and *ampE*, with *ampD-ampE* pair and (c) *ackA* and *pta* with *ackA-pta* pair. We looked for the scores for fragments containing more than one gene that are significantly greater than the inferred sum of score of the constituent genes.

1375 **Supplementary note:**

1376 **Ridge, Lasso, and Elastic Net**

1377 The Ridge, Lasso, and Elastic Net regressions were implemented using the scikit-learn  
1378 python library for machine learning. The regression was done on sparse representation  
1379 of matrix A, without calculation of intercept since fragment scores were normalized (to  
1380 set the median to zero). The regularization parameters were estimated using 3-fold  
1381 cross validation (RidgeCV, LassoCV, and ElasticNetCV classes from the  
1382 sklearn.linear\_model package). The parameters were first estimated for each of 155  
1383 experiments, and then the parameters that deliver the highest R-square across all  
1384 samples were selected as optimal.

1385  
1386 The objective functions to be minimized and optimal regularization parameters for  
1387 Ridge, Lasso, and Elastic Net are described below.

1388  
1389 **Ridge**

1390  
1391 Ridge is  $L_2$  regularization with objective function:

$$\|Ag - f\|_2^2 + \alpha \|g\|_2^2$$

1393  
1394 where  $\alpha$  controls the amount of regularization (shrinkage). The optimal  $\alpha = 1.0$

1395 **Lasso**

1396 Lasso is  $L_1$  regularization with objective function:

$$\|Ag - f\|_2^2 + \alpha \|g\|_1$$

1398  
1399 where  $\alpha$  controls the amount of regularization (shrinkage) and variable selection. The  
1400 optimal  $\alpha = 3.4$

1401  
1402 **Elastic Net**

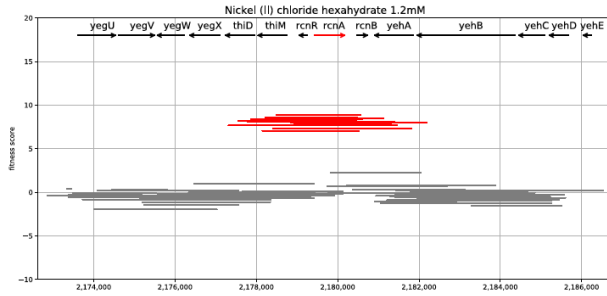
1403 Elastic Net is regularization with linear combination of  $L_1$  and  $L_2$  terms and objective  
1404 function:

$$\|Ag - f\|_2^2 + \alpha \gamma \|g\|_1 + \frac{\alpha(1-\gamma)}{2} \|g\|_2^2$$

1406  
1407 where  $\alpha$  controls the amount of regularization and  $\gamma$  defines the relative contribution of  
1408  $L_1$  and  $L_2$  terms/ The optimal parameters:  $\alpha = 3.6$ ;  $\gamma = 0.7$

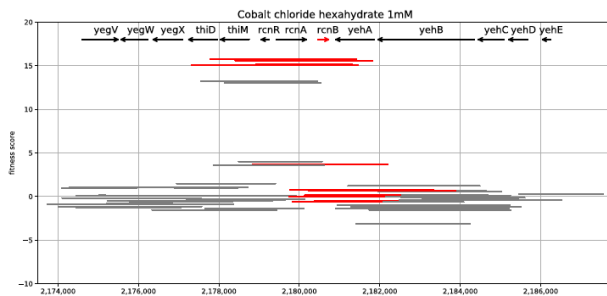
1409 The regression analysis was run using optimal parameters and then manual inspection  
1410 of regression results obtained from all three methods (Ridge, Elastic Net and LASSO)  
1411 was performed for known gene-function associations. We observed that Ridge and  
1412 Elastic Net with optimal parameters tends to significantly underestimate the fitness  
1413 scores for causative genes that expected to have high positive or negative fitness  
1414 scores. This underestimation is caused by shrinkage effect introduced by both  
1415 regularization approaches. At the same time, the LASSO, when used with optimal  
1416 parameters, seems to lack this problem and produces the most accurate scores across  
1417 all three approaches. As an example, this is shown for *rcnA* gene (condition: 1.2 mM  
1418 Nickel) scores calculated from Ridge, Elastic Net and LASSO approaches  
1419 (**Supplementary Fig. 7a**). However, LASSO with optimal parameters still did not solve  
1420 OLS over fitting problem completely, and still gave the unrealistic extreme positive and  
1421 extreme negative scores for neighboring genes (for example, comparison of *rcnB* and  
1422 *yehA*, condition: 1mM Cobalt, **Supplementary Fig. 7bc**). In comparison, NNLS had no  
1423 regularization parameters, and we did not observe over fitting issues.  
1424  
1425

a



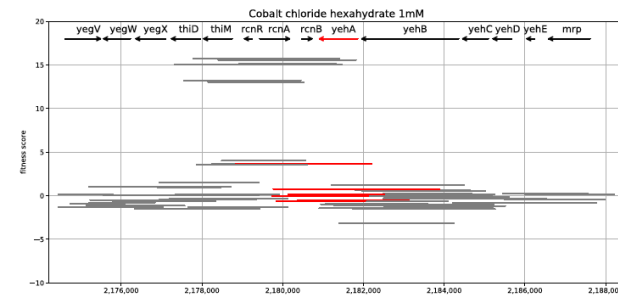
gene_name	rcnA
locus_tag	BW25113_2106
score_nnl5	7.98483
score_ridge	<u>5.87525</u>
score_lasso	7.82251
score_enet	<u>6.73892</u>

b



gene_name	rcnB
locus_tag	BW25113_2107
score_nnl5	1.90233
score_ridge	5.68672
score_lasso	<u>7.47208</u>
score_enet	6.04639

c



gene_name	yehA
locus_tag	BW25113_2108
score_nnl5	0
score_ridge	-5.39295
score_lasso	<u>-7.65625</u>
score_enet	-5.86752

1426

1427

1428 **Supplementary Fig. 7: Gene score estimation approaches:** Example gene scores  
 1429 for (a) *rcnA* (b) *rcnB* and (c) *yehA* showing data over fitting and shrinkage by ridge,  
 1430 lasso and elastic net regularization methods. Left, Dub-seq viewer for fragments  
 1431 covering a specific gene completely (red), compared to partially covering or gene-  
 1432 neighborhood fragments (gray). The gene scores estimated using different methods are  
 1433 shown on right. The gene scores highlighted in blue lines indicate issues of  
 1434 regularization methods (see Supplementary note).

1435

1436