



**HAL**  
open science

## The impact of genetic diversity on gene essentiality within the *E. coli* species

François Rousset, José Cabezas Caballero, Florence Piastra-Facon, Jesús Fernández-Rodríguez, Olivier Clermont, Erick Denamur, Eduardo P. C. Rocha, David Bikard

### ► To cite this version:

François Rousset, José Cabezas Caballero, Florence Piastra-Facon, Jesús Fernández-Rodríguez, Olivier Clermont, et al.. The impact of genetic diversity on gene essentiality within the *E. coli* species. *Nature Microbiology*, 2021, 6 (3), pp.301-312. 10.1038/s41564-020-00839-y . pasteur-03264663

**HAL Id: pasteur-03264663**

**<https://pasteur.hal.science/pasteur-03264663v1>**

Submitted on 18 Jun 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The impact of genetic diversity on gene essentiality within the *E. coli* species

François Rousset<sup>1,2</sup>, José Cabezas Caballero<sup>1</sup>, Florence Piastra-Facon<sup>1</sup>, Jesús Fernández-Rodríguez<sup>3</sup>, Olivier Clermont<sup>4</sup>, Erick Denamur<sup>4,5</sup>, Eduardo P.C. Rocha<sup>6,\*</sup> & David Bikard<sup>1,\*</sup>

- 1- Synthetic Biology, Department of Microbiology, Institut Pasteur, Paris, France
- 2- Sorbonne Université, Collège Doctoral, F-75005 Paris, France
- 3- Eligo Bioscience, Paris, France
- 4- Université de Paris, IAME, INSERM UMR1137, Paris, France
- 5- AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, Paris, France
- 6- Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, 25-28 rue Dr Roux, Paris, 75015, France.

\*To whom correspondence should be addressed: [david.bikard@pasteur.fr](mailto:david.bikard@pasteur.fr) or [eduardo.rocha@pasteur.fr](mailto:eduardo.rocha@pasteur.fr)

## Abstract

Bacteria from the same species can differ widely in their gene content. In *E. coli*, the set of genes shared by all strains, known as the core genome, represents about half the number of genes present in any strain. While recent advances in bacterial genomics have unraveled genes required for fitness in various experimental conditions at the genome scale, most studies have focused on single model strains. As a result, the impact of this genetic diversity on core processes of the bacterial cell remains largely under-investigated. Here, we developed a new CRISPR interference platform for high-throughput gene repression that is compatible with most *E. coli* isolates and closely-related species. We applied it to assess the importance of ~3,400 nearly ubiquitous genes in 3 growth conditions in 18 representative *E. coli* strains spanning most common phylogroups and lifestyles of the species. Our screens revealed extensive variations in gene essentiality between strains and conditions. Investigation of the genetic determinants for these variations highlighted the importance of epistatic interactions with mobile genetic elements. In particular, we showed how prophage-encoded defense systems against phage infection can trigger the essentiality of persistent genes that are usually nonessential. This study provides new insights into the evolvability of gene essentiality and argues for the importance of studying various isolates from the same species under diverse conditions.

## 1 Introduction

2 Essential genes can be defined as genes required for the reproduction of an organism<sup>1</sup>. They are  
3 thought to be rarely lost because of their essentiality and have a lower substitution rate than other  
4 genes<sup>2,3</sup>. However, previous work showed that closely-related taxa have different essential genes<sup>4-10</sup>. In *E.*  
5 *coli*, the Keio collection<sup>11,12</sup> and transposon-sequencing methods<sup>13-15</sup> have enabled the determination of  
6 genes required for growth in various conditions, but were mostly limited to the laboratory-evolved model  
7 strain K-12. This strain is not representative of the broad diversity of the *E. coli* species which is  
8 characterized by an open pangenome with high rates of horizontal gene transfer (HGT)<sup>16-18</sup>. This broad  
9 genetic diversity results in the adaptation of *E. coli* strains to multiple ecological niches and lifestyles: *E. coli*  
10 can be found in the environment as well as in association with humans and animals where it can behave as  
11 a gut commensal or as an opportunistic intestinal and extra-intestinal pathogen<sup>19,20</sup>. A few studies have  
12 used transposon-sequencing to determine the genetic requirements of clinical *E. coli* isolates for *in vitro*  
13 growth or colonization of animal models<sup>21-25</sup>. This showed that clinical strains associated with different  
14 pathologies require different genes for colonization and virulence. Although these findings represent an  
15 important insight into the mechanisms of infection, a direct comparison of growth requirements of *E. coli*  
16 strains is still lacking. In particular, the broad genetic diversity of *E. coli* provides the opportunity to assess  
17 how the genetic background influences gene essentiality.

18 Several hypotheses could explain why genetic diversity may impact gene essentiality. A gene that is  
19 essential in a strain might be dispensable in another strain if the latter carries a homolog or an analog that  
20 performs the same function. In this situation, the pair of genes is known as synthetic lethal. Previous  
21 studies also showed that the loss of some essential genes can be compensated by the overexpression of  
22 genes carrying a different function<sup>26,27</sup>. Another example is the case of prophage repressors and antitoxins<sup>28</sup>  
23 which typically belong to the accessory genome and are only essential when the cognate prophage or toxin  
24 is also present. It remains unclear if there are significant variations in the essential character of core genes  
25 across the *E. coli* species. A recent investigation of a panel of 9 *Pseudomonas aeruginosa* strains showed  
26 that gene essentiality indeed varies between strains<sup>7</sup>, but the underlying mechanisms and the relevance of  
27 these findings to other bacterial species remain to be investigated.

28 To tackle this question, we turned to CRISPR interference (CRISPRi). This method is based on the  
29 catalytically-inactivated variant of Cas9, dCas9, which can be directed by a single-guide RNA (sgRNA) to bind  
30 a target gene and silence its expression<sup>29-31</sup>. Using sgRNA libraries, pooled CRISPRi screens were recently  
31 developed in bacteria to investigate the contribution of each gene to fitness by monitoring the fold-change  
32 in sgRNA abundance during growth using deep sequencing<sup>32</sup>. Since a strong contribution to fitness is  
33 generally a good proxy for gene essentiality, such screens were used to identify essential genes in *E. coli*<sup>33-37</sup>  
34 and in a few other bacterial species<sup>38-40</sup>. Here, we developed an easy-to-use CRISPRi screening platform  
35 that is compatible with most *E. coli* isolates and closely-related *Enterobacteriaceae* species. We then  
36 designed a compact sgRNA library targeting the *E. coli* core genome in order to compare the essentiality of  
37 core genes in different genetic backgrounds and growth conditions. Our results reveal how the essentiality  
38 of core genes can substantially vary at the strain level. Further investigation of the underlying mechanisms  
39 showed that HGT and gene loss events can modulate the essentiality of core genes.

## 40 Results

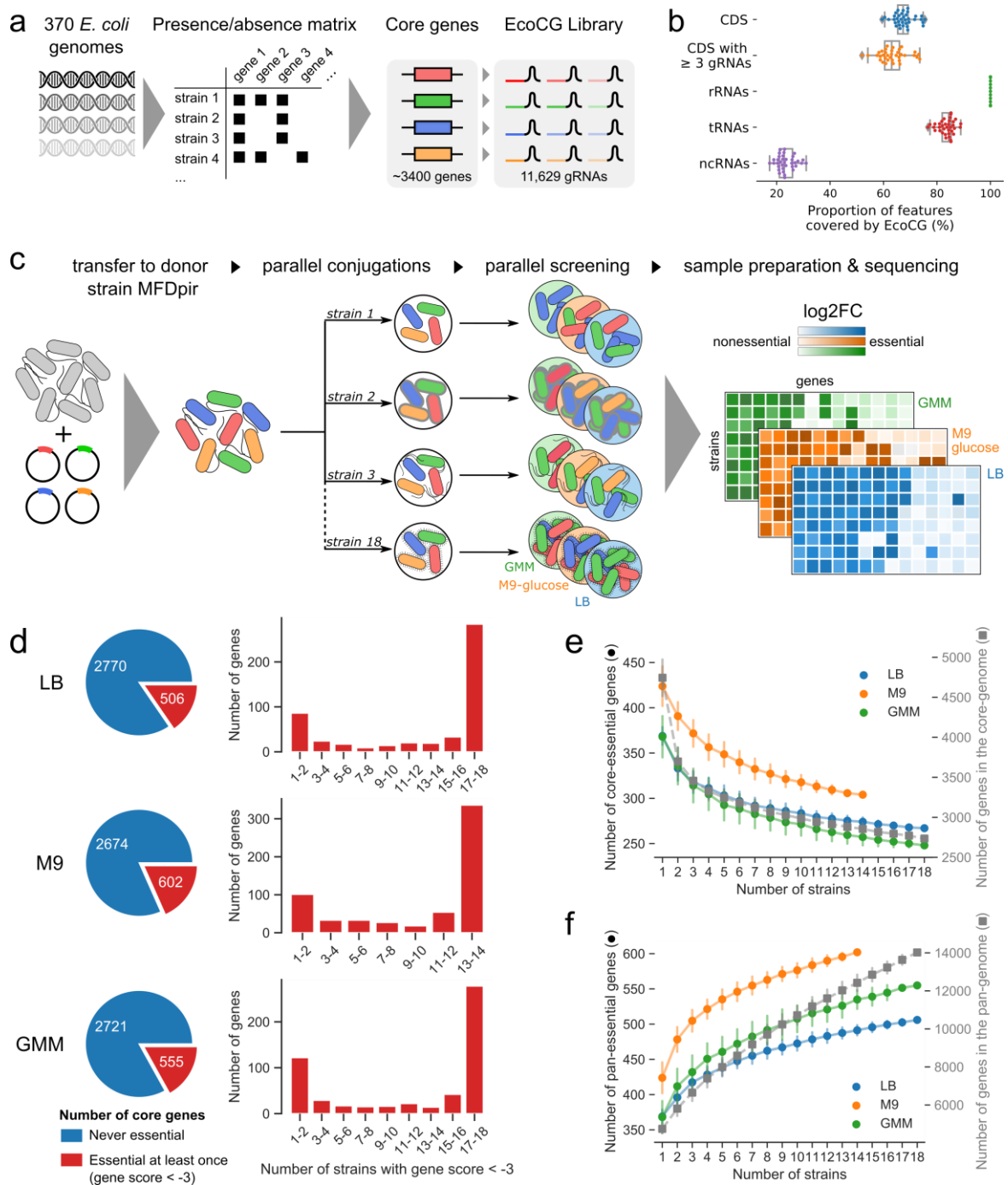
### 41 A compact sgRNA library targeting ~3,400 nearly-ubiquitous genes from *E. coli*

42 We first designed an easy-to-use single plasmid vector for CRISPRi called pFR56, comprising a  
43 constitutively expressed sgRNA, a *dCas9* expression cassette controlled by a DAPG-inducible PhIF  
44 promoter<sup>41</sup> and an RP4 origin to enable transfer by conjugation. In order to ensure plasmid stability in most  
45 strains, protein-coding sequences from pFR56 were recoded to avoid most restriction sites that are  
46 recognized by the restriction-modification systems of *E. coli*<sup>42</sup>. We further optimized *dCas9* expression level  
47 to ensure that its expression was non-toxic (**Supplementary Fig. 1a**). pFR56 achieved a high conjugation  
48 rate and was stable for >24 generations without antibiotic selection in various strains from species  
49 belonging to the *Escherichia*, *Klebsiella* and *Citrobacter* genera (**Supplementary Fig. 1b-c**). We confirmed  
50 efficient *dCas9*-mediated repression in these strains by measuring growth inhibition when targeting the  
51 essential gene *rpsL* (**Supplementary Fig. 1d**). This demonstrates the usefulness of our CRISPRi system in a  
52 broad range of *E. coli* isolates and in closely related species.

53 While most studies in *E. coli* rely on lab-evolved derivatives of strain K-12, we aimed at investigating  
54 gene essentiality in the *E. coli* species as a whole. The size of the pangenome makes it impossible to target  
55 all genes from the species. Instead, focusing on core genes enables a direct comparison of the same genes  
56 under different genetic backgrounds. We analyzed 370 complete *E. coli* genome sequences and identified  
57 3380 protein-coding genes present in > 90 % of genomes (**Fig. 1a**). We then selected 3-4 sgRNAs per gene  
58 by favoring targets that are conserved across strains while maximizing predicted on-target activity<sup>36</sup>,  
59 minimizing off-target activity and avoiding toxic seed sequences<sup>33</sup> (**see Methods**). The resulting *E. coli* core  
60 genome (EcoCG) library comprises 11,629 sgRNAs targeting ~60-80% of the protein-coding genes of any *E.*  
61 *coli* strain as well as 100% of rRNAs, 75-85% of tRNAs and 15-25% of annotated ncRNAs (**Fig. 1b**). The EcoCG  
62 library was cloned onto pFR56 and transferred to K-12 MG1655 by conjugation in order to evaluate its  
63 performance in the prediction of essential genes during growth in LB (**Supplementary Fig. 1e**). A screen  
64 performed using this library predicted essential genes better than a previous randomly-designed genome-  
65 wide library<sup>34</sup> (AUC = 0.979 vs 0.963, Delongs' test<sup>43</sup>  $Z = -2.4754$ ,  $p = 0.013$ ) despite being much smaller (~9  
66 vs 3.4 sgRNAs per gene on average) (**Supplementary Fig. 1g-h**), highlighting the benefits of an improved  
67 design.

### 68 Distribution of gene essentiality in an *E. coli* strain panel

69 We selected a panel of 18 *E. coli* natural isolates spanning most common *E. coli* phylogroups (A, B1, B2, D, E  
70 and F) and lifestyles in order to compare the essentiality of their conserved genes (**Supplementary Table 1,**  
71 **see Methods**). This panel includes the lab-derived strain K-12 MG1655, environmental isolates (E1114,  
72 E1167 and E101), commensals from humans (HS) and other mammals (M114, ROAR8, TA054, TA249, TA280  
73 and TA447), an intestinal pathogen associated with Crohn's disease (41-1Ti9) and extra-intestinal  
74 pathogens isolated from blood, lungs, urine and cerebrospinal fluid from humans and poultry (H120,  
75 JJ1886, APEC O1, S88, CFT073, UTI89). In order to compare genetic requirements for growth in various  
76 experimental contexts, we performed CRISPRi screens with each strain in two biological replicates during  
77 aerobic growth in LB or in minimal M9-glucose medium (M9), as well as during anaerobic growth in gut  
78 microbiota medium (GMM)<sup>44</sup> (**Supplementary Table 2, Fig. 1c**), yielding a total of 100 CRISPRi screens on  
79 ~3400 genes (four strains were discarded in M9 due to insufficient growth). Thanks to the small size of the  
80 EcoCG library, all screening results could be obtained from a single Illumina NextSeq 500 run, representing  
81 a cost of less than 20€ per sample. Biological replicates achieved a very high reproducibility (median  
82 Pearson's  $r = 0.988$ ), demonstrating the robustness of the method (**Supplementary Fig. 2a**). For each



83

84 **Figure 1 | Distribution of fitness defects after CRISPRi screening in 18 *E. coli* strains and 3 media with the EcoCG library.** **a**,  
 85 Starting from 370 complete *E. coli* genomes, a gene presence/absence matrix was computed to deduce 3,380 protein-coding genes  
 86 that are present in > 90% *E. coli* strains. For each gene, 3 or 4 sgRNAs were selected based on the proportion of targeted strains and  
 87 on the predicted off-target activity, efficiency and bad-seed effect. We also added sgRNAs targeting rRNAs, tRNAs and widespread  
 88 ncRNAs (see Methods), yielding the EcoCG library comprising 11,629 sgRNAs. **b**, The EcoCG library was mapped to the genome of  
 89 42 *E. coli* strains. On average, it targets 67.7% of the protein-coding gene content (with 100% nucleotide identity), and 63.4 % with  
 90 at least 3 sgRNAs. **c**, The MFDpir conjugation strain<sup>45</sup> was used to transfer the EcoCG library to a panel of 18 *E. coli* isolates. Each  
 91 strain was then grown for 20 generations with dCas9 induction in aerobic conditions in LB and M9-glucose medium and in  
 92 anaerobic condition in gut microbiota medium (GMM). Log<sub>2</sub>FC and gene score values were computed (see Methods). Only 14  
 93 strains were screened in M9-glucose due to poor growth of 4 strains. **d**, For each medium, we selected core genes whose  
 94 repression induces a fitness defect in at least one strain (gene score < -3) (left) and reported the number of strains where this  
 95 defect can be seen (right). **(e-f)** Evolution of the number of core genes that are essential in all strains **(e)** or in at least one strain **(f)**  
 96 as a function of the number of selected strains (circle markers). The error bars indicate the standard-deviation of up to 250 random  
 97 permutations. The grey dashed curves represent the size of the core genome **(e)** or the size of the pangenome **(f)** (Square markers)  
 98 with the scale shown on the right.

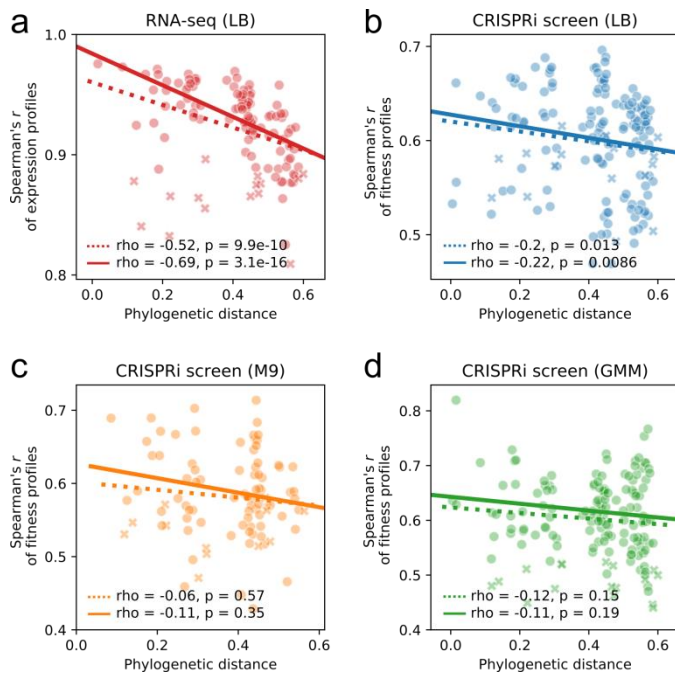
99 screen, gene scores were calculated as the median log<sub>2</sub>FC of sgRNAs targeting the gene (**Supplementary**  
100 **Tables 3-4**), resulting in a gene-strain scoring matrix for each of the three tested media. In the following  
101 analyses, we considered genes with a score lower than -3 as essential in a given strain and condition. This  
102 stringent threshold recovers 86.3% of known essential genes from K-12 in LB<sup>13</sup> with a false-positive rate of  
103 only 2.7% (**Supplementary Fig. 1g**), mostly due to expected polar effects<sup>34</sup> (**Supplementary Fig. 1i**). Note  
104 that this relaxed definition of essentiality includes all genes whose repression leads to a major fitness  
105 defect.

106 We first investigated the overall genetic requirements of the *E. coli* species for growth in each  
107 medium, recapitulating previous findings and refining our knowledge of common requirements across  
108 growth conditions (**Supplementary Results**). We then explored how many genes are essential in at least  
109 one strain and how frequently these genes are essential. We found many more genes that are essential in  
110 at least one strain than genes that are essential in all tested strains (506 vs 266 in LB, 602 vs 304 in M9 and  
111 555 vs 248 in GMM) (**Fig. 1d, Supplementary Table 5**). Most essential genes are either essential in most  
112 strains or in a small number of them (**Fig. 1d**). These results show that the essentiality of core genes varies  
113 substantially at the strain level. We can tentatively use this data to define a core-essential genome (i.e.  
114 genes that are virtually essential in all strains of the species) and a pan-essential genome (i.e. genes that are  
115 essential in at least one strain of the species). We performed a rarefaction analysis by computing the core-  
116 essential genome and pan-essential genome for various sets of strains. Interestingly, the size of the core  
117 genome and the size of the core-essential genome converge at a similar pace (**Fig. 1e**). As a result, the  
118 fraction of the core genome that is essential in all strains is roughly independent from the number of  
119 strains under consideration (e.g. ~9-10% of the core genome in LB) (**Supplementary Fig. 4a**). The set of core  
120 genes that are essential in at least one strain keeps increasing with the addition of new strains (**Fig. 1f,**  
121 **Supplementary Fig. 4b**), showing that our results probably only reveal a fraction of the existing differences  
122 at the species level and that a significant part of the nonessential core genome is likely to become essential  
123 in certain genetic backgrounds.

## 124 **The impact of phylogeny on gene expression and essentiality**

125 Since gene essentiality has been linked to a higher gene expression level<sup>3</sup>, we wondered to what  
126 extent changes in gene essentiality are reflected by changes in gene expression level. We generated RNA-  
127 sequencing (RNA-seq) data for 16 strains during growth in exponential phase in LB and compared the  
128 expression of core genes (**Supplementary Table 6, see Methods**). As previously observed<sup>46</sup>, expression level  
129 and essentiality were correlated, with a higher expression level for essential genes (**Supplementary Fig. 5a**).  
130 We wondered whether this was also the case for genes whose essentiality varies, i.e. if a shift in essentiality  
131 is associated with a shift in expression. We selected 87 genes that were variably essential between the 16  
132 strains assayed in RNA-seq experiments (**see Methods**). Considering all strains together, these “variably  
133 essential” genes tend to be more expressed than genes that are never essential but less expressed than  
134 genes that are always essential (**Supplementary Fig. 5b**). When considering each “variably essential” gene  
135 individually, we found no correlation between CRISPRi fitness and gene expression level across the 16  
136 strains (**Supplementary Fig. 5c**), suggesting that a shift in essentiality is not associated with a shift in  
137 expression level.

138 We then investigated the importance of phylogeny in the variations in gene expression and  
139 essentiality. We observed a strong correlation between the phylogenetic distance of pairs of strains and  
140 their similarity in gene expression profile (Spearman’s rho = -0.52 p < 10<sup>-9</sup>), i.e. closely-related strains have  
141 more similar expression profiles (**Fig. 2a**). Interestingly, K-12 MG1655 seems to be an outlier and discarding  
142 it from this analysis markedly improved the correlation (rho = -0.69, p < 10<sup>-15</sup>) (**Fig. 2a and Supplementary**



**Figure 2 | The impact of phylogeny on gene expression and essentiality.** Regressions show the relationship between the phylogenetic distance of pairs of strains and the Spearman correlation of their gene expression profiles during exponential growth in LB (a) or with the Spearman correlation of their CRISPRi fitness profiles in LB (b), M9 (c) or GMM (d). Each dot represents a pair of strains and data points corresponding to K-12 MG1655 are shown with a cross marker. The dotted line represents the regression considering all strains while the solid line represents the regression when excluding K-12 MG1655. Spearman rho coefficients are shown for each regression. a, RNA-seq data was obtained on 16 strains, representing 120 pairs (105 when excluding K-12 MG1655). (b,d) CRISPRi screening data in LB and GMM was obtained on 18 strains, representing 153 pairs (136 when excluding K-12 MG1655). c, CRISPRi screening data in M9 was obtained on 14 strains, representing 91 pairs (78 when excluding K-12 MG1655).

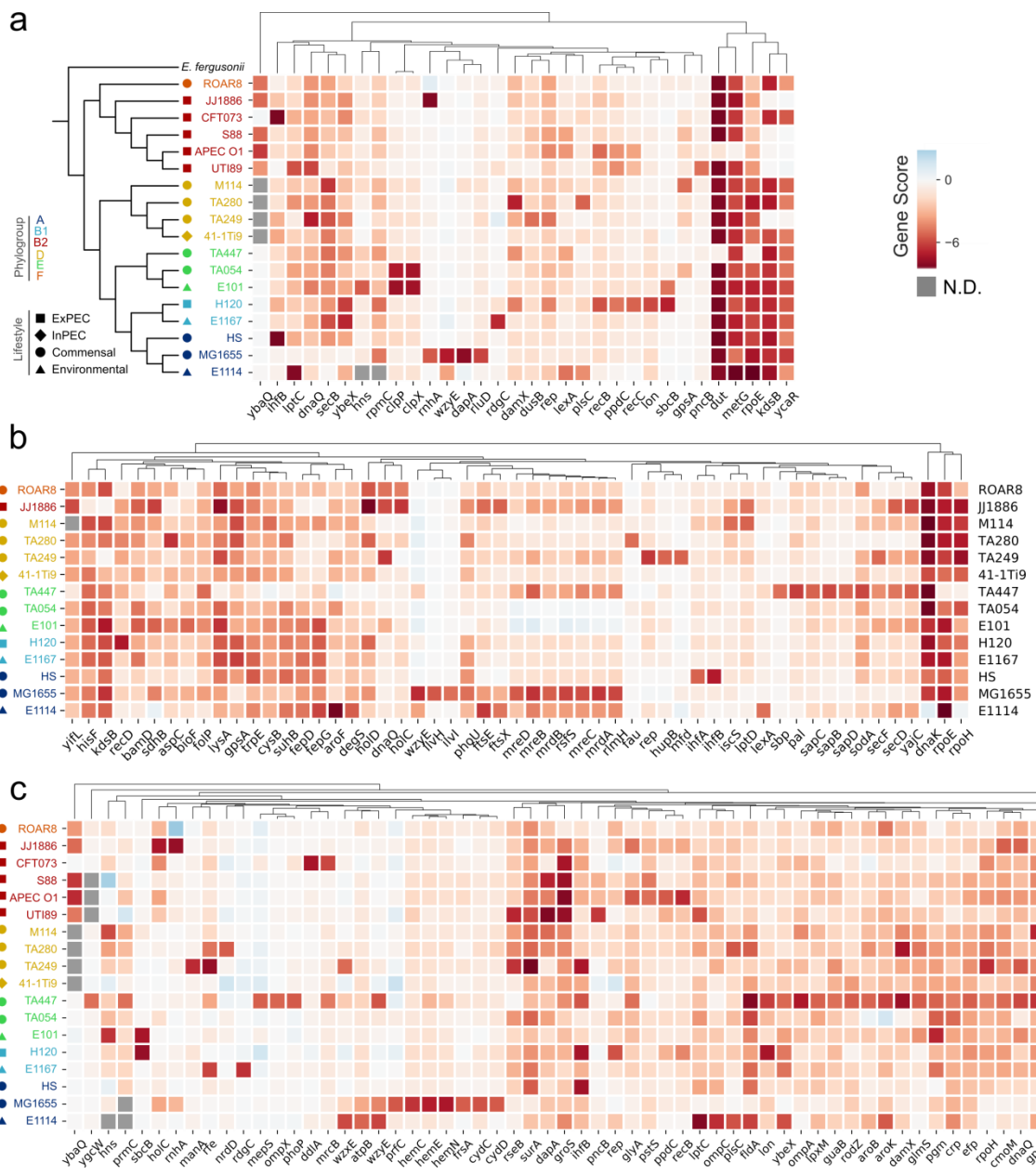
143 **Fig. 6a**), possibly because of mutations acquired during laboratory evolution. We then conducted the same  
 144 analysis with the CRISPRi fitness profiles. The correlation was weak in LB ( $\rho \sim 0.2$ ,  $p \sim 0.01$ ) and was not  
 145 significant in M9 and in GMM regardless of the inclusion of K-12 MG1655 (**Fig. 2b-d, Supplementary Fig.**  
 146 **7a**), showing that the evolutionary distance is a poor predictor of the similarity in fitness profiles. We then  
 147 wondered whether pairs of strains that share an essential gene tend to be more closely related than pairs  
 148 for which the gene is differentially essential. This is indeed the case for a handful of genes whose  
 149 contribution to fitness changes in some clades (see for instance the cases of *kdsB* and *ybaQ* below), but no  
 150 signal could be seen for the majority of “variably essential” genes (**Supplementary Fig. 7b**). These results  
 151 suggest that while changes in the expression level of core genes are strongly linked to vertical inheritance,  
 152 phylogeny has a weaker influence on the contribution of core genes to fitness.

### 153 **Homologs, analogs and functional redundancy**

154 In order to investigate the genetic basis for differences in essentiality, we focused on cases where  
 155 the difference is large by selecting genes whose repression induces a very strong fitness defect (gene score  
 156  $< -5$ ) in at least one strain, while having no effect (gene score  $> -1$ ) in at least one strain (**see Methods**). This  
 157 resulted in 32 protein-coding genes in LB (**Fig. 3a**), 55 in M9 (**Fig. 3b**) and 66 in GMM (**Fig. 3c**), representing  
 158 a total of 120 unique genes which displayed very distinct degrees of essentiality across strains  
 159 (**Supplementary Fig. 8**). We then aimed at determining the genetic mechanisms explaining these  
 160 differences.

161 We first investigated whether some effects could be linked to the presence of functional homologs  
 162 making an essential gene dispensable in some strains. Our screens showed that all strains where the *ycaR*-  
 163 *kdsB* transcriptional unit (expressing the essential CMP-KDO synthetase KdsB) is dispensable carry another  
 164 CMP-KDO synthetase gene, *kpsU*, whose product shares 46% of identity with KdsB. Simultaneous  
 165 repression of both *kdsB* and *kpsU* induced a strong fitness defect in strains that are resistant to *kdsB*  
 166 knockdown (**Supplementary Fig. 9**). Another example is the case of *metG* (methionine-tRNA ligase) which  
 167 can be explained by a gene duplication event (**Supplementary Results**).





168

169 **Figure 3 | Extensive differences in gene essentiality within *E. coli* core genes.** For each growth condition, we selected genes whose  
 170 repression produces a strong fitness defect in at least one strain (gene score < -5) while producing no effect in at least one strain  
 171 (gene score > -1) (see Methods). Panels (a), (b) and (c) show a heatmap of the selected genes in LB, M9 and GMM respectively.  
 172 Genes were clustered by Euclidian distance. Grey squares correspond to genes and strains where no gene score is available (N.D.,  
 173 non-determined). The phylogeny and origin of the strains are shown in panel a.

174 We attempted to assess how frequently the existence of homologs makes an essential gene  
 175 dispensable. Overall, the strains we tested carry 10 to 17 homologs (>40% identity, median = 13.5) of core  
 176 genes that are essential in *E. coli* K-12 in LB<sup>13</sup>, including 3 to 7 (median = 4) with >60% identity. This shows  
 177 that homologs of essential genes are relatively frequent. We might therefore expect more cases of  
 178 essential genes that become nonessential in some strains because of genetic redundancy. However, this  
 179 seems to be the case only for the genes detailed above. As an example, *nrdA* and *nrdB* remain essential in  
 180 APEC O1 and TA447 despite the presence of two homologs whose products share 63% and 60% of identity  
 181 to NrdA and NrdB respectively. RNA-seq data showed that these homologs are poorly expressed in our  
 182 experimental conditions (< 2% of the expression level of *nrdA* and *nrdB*) which likely explains their inability  
 183 to rescue the repression of *nrdAB*. In addition, putative homologs of essential genes may not be  
 184 functionally redundant since gene homology does not necessarily imply functional redundancy<sup>47</sup>.



185 Altogether, our data shows that with a few exceptions documented here, the presence of a homolog does  
186 not typically provide functional redundancy.

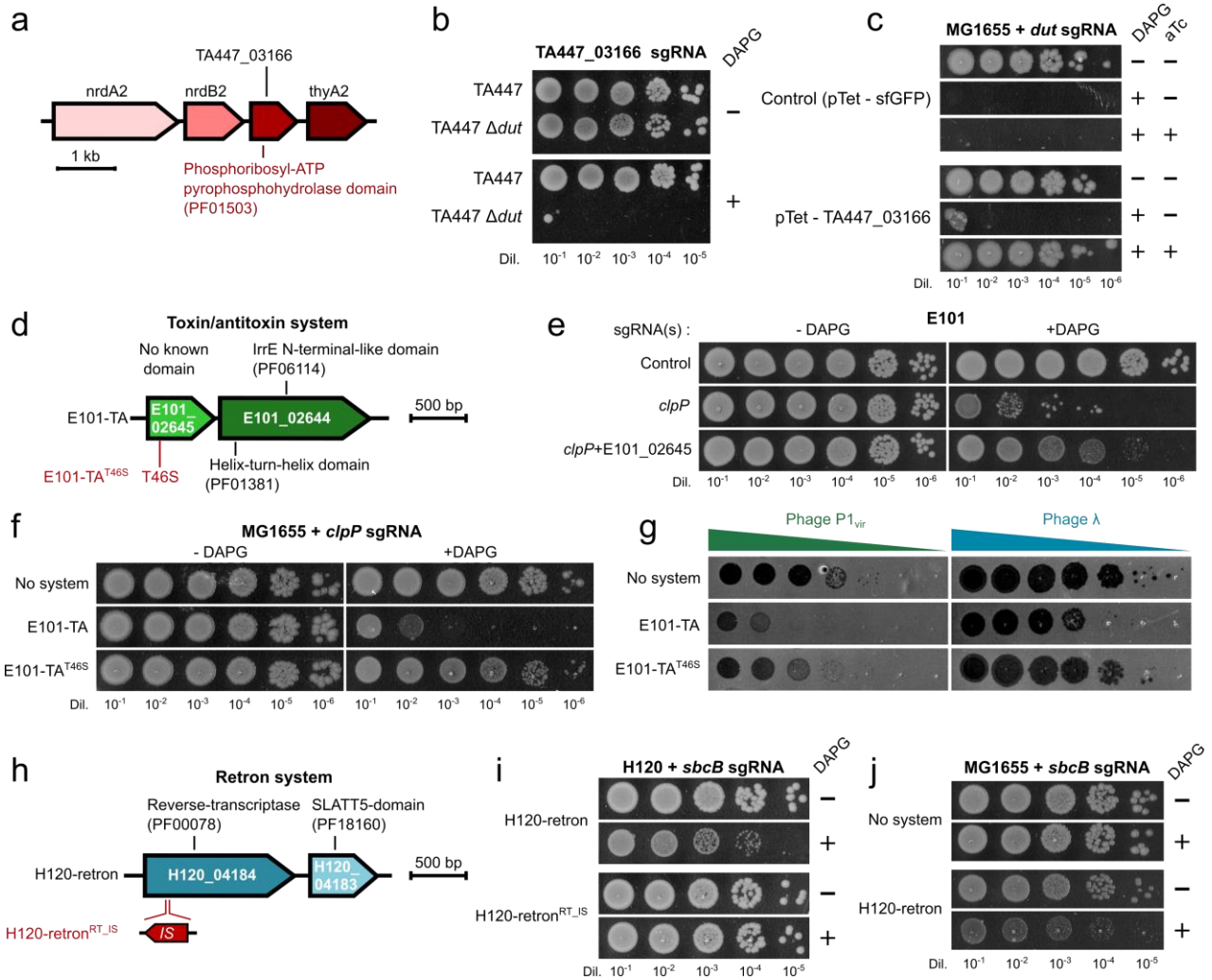
187 An essential gene may also become dispensable because of the acquisition of genes of analogous  
188 function by HGT. The product of *dut* (dUTPase) hydrolyses dUTP into dUMP in order to avoid its  
189 incorporation into DNA<sup>48</sup>. This gene is essential except in APEC O1 and TA447 (**Fig. 3**), but we did not find  
190 any sequence homolog of Dut in these two genomes. We successfully built a TA447 $\Delta$ *dut* strain and verified  
191 the absence of compensatory mutations, confirming that *dut* is indeed nonessential in this strain. Further  
192 investigation showed that the plasmid that is shared between APEC O1 and TA447 contains nucleotide  
193 biosynthesis genes (**Fig. 4a**). In particular, a hypothetical protein (TA447\_03166, accession:  
194 WP\_085453089.1) has an ATP-pyrophosphohydrolase-like domain and shows structural homology to a  
195 MazG-like protein from *Deinococcus radiodurans* and other NTP-pyrophosphatases. This gene became  
196 essential in TA447 when *dut* was deleted (**Fig. 4b**), and rescued the growth of K-12 MG1655 when *dut* was  
197 repressed (**Fig. 4c**). Interestingly, *dut* repression induced a fitness defect in TA447 in GMM (**Supplementary**  
198 **Fig. 10a**), a phenomenon likely linked to the lower expression level of TA447\_03166 in GMM  
199 (**Supplementary Fig. 10b**). Taken together, these results suggest that TA447\_03166 encodes a dUTPase that  
200 is functionally redundant with Dut. In this case, compensation is provided by a functional analog sharing no  
201 sequence homology with the essential protein.

## 202 Mapping the genetic determinants of strain-specific essentiality

203 In contrast with *kdsB*, *metG* and *dut*, most “variably essential” genes are essential in very few  
204 strains (**Fig. 1b** and **Fig. 3**). This suggests that core genes can frequently become essential in a few genetic  
205 backgrounds. Most strains (72%, 13/18) had at least one strain-specific or near-specific (in  $\leq 2$  strains)  
206 essential gene in at least one medium and K-12-specific essential genes were the most abundant. We first  
207 looked in the literature for evidence explaining such differences. For instance, *rluD* (23S rRNA  
208 pseudouridine synthase) is known to be essential in K-12 because of an epistatic interaction with a  
209 mutation acquired in *prfB* during laboratory evolution<sup>49</sup>. Other cases of epistatic interactions in K-12 include  
210 the acetohydroxy acid synthase III (AHAS III) encoded by *ilvHI* in M9 (**Fig. 3b**). An isozyme encoded by *ilvGM*  
211 can perform the same essential reaction in other strains but is disrupted in K-12 by a frame-shift, making  
212 AHAS III essential. In addition, we also found that *ybaQ*, a transcriptional regulator of unknown function, is  
213 actually an antitoxin that is essential in B2/F strains that carry *higB-1*, a clade-specific toxin upstream *ybaQ*  
214 (**Supplementary Results, Supplementary Fig. 11**).

215 In order to better understand why some genes become essential under certain genetic  
216 backgrounds, we then set up a pipeline to isolate mutants that suppress strain-specific essentiality and  
217 identified the responsible mutations by whole-genome sequencing (**see Methods**). In this way, we isolated  
218 mutants of the environmental strain E101 that suppress the requirement for the AAA+ protease ClpP which  
219 is essential both in E101 and TA054 (**Fig. 3**). Whole-genome sequencing of 4 suppressor mutants revealed  
220 that a prophage also present in TA054 was excised in 3 clones and that a hypothetical protein (E101\_02645,  
221 accession: WP\_001179380.1) from the same prophage had a non-synonymous (T46S) mutation in the  
222 remaining clone (**Fig. 4d**). This protein has no predicted domain but seems to be the toxic component of a  
223 toxin-antitoxin (TA) module with the downstream protein E101\_02644 (accession: WP\_000481765.1) which  
224 has a helix-turn-helix and a ImaA/IrrE family metallo-endopeptidase domain (these proteins respectively  
225 share 33% and 42% of identity with the two components of a putative TA module from *Vibrio mimicus*<sup>50</sup>).  
226 Targeting E101\_02645 with dCas9 partially rescued the toxicity associated with *clpP* repression in E101 (**Fig.**  
227 **4e**), while this system, but not the T46S variant, made *clpP* essential in K-12 once heterologously expressed  
228 (**Fig. 4f**). Using a larger panel of 48 *E. coli* strains, we identified 6 strains from various phylogroups where

229 *clpP* is essential (CIP61.11, E101, E2348/69, H263, LF82 and TA054), all of which share a distant homolog of  
 230 this system. In an atypical TA system from *Caulobacter crescentus*, the antitoxin is required for ClpXP-  
 231 mediated degradation of the toxin<sup>51</sup>. Although the two systems are not related, a similar mechanism could  
 232 explain our results. Since TA systems have been involved in anti-phage defense<sup>52</sup>, we measured the  
 233 susceptibility of K-12 MG1655 to infection by different phages when expressing this system. The wild-type  
 234 system reduced the efficiency of plaquing of phage P1vir by >100-fold and of phage  $\lambda$  by >10-fold while the  
 235 T46S mutation almost completely abolished resistance (Fig. 4g), showing that this TA system participates in  
 236 bacterial immunity.



237

238 **Figure 4 | Genes encoded on mobile genetic elements can modulate the essentiality of core genes.** (a-c) A plasmid-borne  
 239 dUTPase makes *dut* nonessential in strains TA447 and APECO1. **a**, A region encoding nucleotide biosynthesis genes was identified in  
 240 TA447 and contains a gene encoding a hypothetical protein (TA447\_03166) with a Phosphoribosyl-ATP pyrophosphohydrolase  
 241 domain (pfam 01503). **b**, Drop assay showing that dCas9 induction with DAPG in the presence of a sgRNA targeting TA447\_03166 is  
 242 lethal when *dut* is deleted from TA447. **c**, TA447\_03166 was cloned and expressed from an aTc-inducible pTet promoter in K-12  
 243 MG1655. dCas9-mediated silencing of *dut* has no effect when this protein is expressed. (**d-g**) A toxin/antitoxin (TA) system involved  
 244 in phage defense makes *clpXP* essential. **d**, A suppressor of *clpP* essentiality in E101 had a mutation (T46S) in the putative toxin of a  
 245 TA module comprising E101\_02645 and E101\_02644. **e**, Spot assay with or without dCas9 induction with DAPG in the presence of a  
 246 sgRNA targeting *clpP*, E101\_02645 or both simultaneously. **f**, The TA modules from E101 or from the suppressor mutant (T46S)  
 247 were transferred to K-12 MG1655 and the effect of *clpP* knockdown was measured by a spot assay in the presence or absence of  
 248 dCas9 induction. **g**, Sensitivity to phages P1vir and lambda was measured in MG1655 carrying an control vector or a vector  
 249 expressing the wild-type or mutated (T46S) TA system from E101. (**h-j**) A retron system makes *sbcB* essential. **h**, Suppressors of  
 250 *sbcB* essentiality in H120 has an insertion element in the reverse-transcriptase of a retron system encoded by H120\_04184 and  
 251 H120\_04183. **i**, Spot assays showing the sensitivity of wild-type or mutant H120 to *sbcB* knockdown. **j**, Once transferred to K-12  
 252 MG1655, this system induces a fitness defect when *sbcB* is repressed.

253 Using the same strategy, we isolated mutants of the uropathogenic strain H120 that can grow in  
254 the presence of a sgRNA targeting *sbcB* (exodeoxyribonuclease I) which is also essential in E101 (**Fig. 3**).  
255 Two suppressors had an insertion element in a prophage-encoded protein (H120\_04184, accession:  
256 WP\_000344414.1) with a reverse-transcriptase domain (pfam 00078) that belongs to a retron system  
257 (H120-retron) (**Fig. 4h-i**). The accessory gene of the system (H120\_04183, accession: WP\_001352776.1)  
258 contains transmembrane helices and a SLATT domain (pfam 18160) that is predicted to function as a pore-  
259 forming effector initiating cell suicide<sup>53</sup>. Heterologous expression of this system in K-12 MG1655 induced a  
260 fitness defect when *sbcB* was repressed (**Fig. 4j**). Importantly, retrons were recently found to form a new  
261 type of TA working as abortive infection systems<sup>54-57</sup>. In a described example, a retron guards RecBCD  
262 function in the cell and triggers cell suicide when RecBCD is inhibited by an incoming phage<sup>56</sup>. We can  
263 hypothesize that the H120-retron also provides resistance to yet unknown phages by guarding the function  
264 of SbcB.

265 In another interesting example, we isolated suppressor mutants of *rnhA* essentiality in strain JJ1886  
266 which mapped to *rnIA*, the toxin of the RnIAB TA system involved in the defense against phage T4  
267 (**Supplementary Results**). These findings show that horizontally-transferred genetic elements involved in  
268 bacterial immunity can trigger the essentiality of core genes.

## 269 Discussion

270 Instead of a binary trait, gene essentiality can be considered as an extreme fitness defect within a range of  
271 continuous values associated with gene disruption. Here, we highlight how variations in environmental  
272 conditions and genetic backgrounds may modulate the fitness defect of a mutant, or even make the  
273 mutation neutral, extending recent results in bacteria and in yeast<sup>4-10,14</sup>. Our screening methodology, by  
274 effectively looking at fitness effects, provides useful data to query how natural selection of genes is  
275 impacted by these two factors. By investigating a collection of strains representative of the *E. coli* genetic  
276 diversity, we could obtain a rich dataset revealing features of the evolution of gene essentiality which could  
277 not be obtained from previous work on the model strain K-12. We show that the number of pan-essential  
278 genes keeps increasing when adding new strains and that variably essential genes are different between  
279 growth conditions. Therefore, future studies including more strains and growth conditions should further  
280 emphasize the broad impact of genetic diversity on gene essentiality. Note that we do not report strain-  
281 specific essential genes outside the core genome such as antitoxins and phage repressors and we thus  
282 underestimate the actual size of the pan-essential genome.

283 Our investigations showed that gene loss and accretion by horizontal transfer had a key role in  
284 providing genetic backgrounds that explain the observed variations in essentiality (**Supplementary Table**  
285 **10**). In the few cases detailed here, the accessory genes that modulate core gene essentiality were acquired  
286 in mobile genetic elements whose residence time is usually short<sup>18</sup>. As a result, our data shows the ability of  
287 some core genes to become essential on a recurrent basis after the acquisition of new genes by HGT. This  
288 was surprisingly the case of several phage defense systems. This phenomenon could favor the evolutionary  
289 conservation of these core genes despite their dispensability in most conditions. While our investigation of  
290 genetic determinants might be biased towards genes with drastic changes in essentiality status, our work  
291 suggests that HGT is a major contributor to changes in gene essentiality. Since this phenomenon is likely  
292 linked to the high rate of HGT in *E. coli*, we could expect extensive differences in gene essentiality in any  
293 species with an open pan-genome.

294 Beyond these findings, we have only analyzed in detail a few of the more salient features of our  
295 dataset. More work needs to be done, for instance to better understand the various genetic requirements  
296 of *E. coli* strains in different growth media, or to investigate the fitness effects associated with the  
297 depletion of the small RNAs that are included as targets in our library. The data we provide in this study  
298 should also prove useful to other fields concerned with gene essentiality. For example, the definition of a  
299 core-essential genome for the *E. coli* species might foster current efforts in genome reduction, while the  
300 strain-specificity of some essential genes could pave the way for the design of antimicrobials that  
301 selectively remove specific bacterial strains from a community. Altogether, our study highlights the  
302 importance of studying the contribution of genes to fitness in many strains and environments, which is now  
303 made possible by recent advances in bacterial genomics.

## 304 **Methods**

### 305 **Bacterial cultivation**

306 Unless stated otherwise, lysogeny broth (LB) broth was used as a liquid medium and LB + 1.5 % agar as a  
307 solid medium. Kanamycin (Kan) was used at 50 µg / mL, chloramphenicol (Cm) was used at 20 µg/mL,  
308 erythromycin (Erm) was used at 200 µg/mL and carbenicillin was used at 100 µg / mL. dCas9 expression  
309 was induced with 50 µM DAPG (Acros Organics). The composition of media used for screening is described  
310 in **Supplementary Table 2**. *E. coli* K-12 MG1655 was used for cloning and MFDpir was used as a donor strain  
311 for plasmid transfer by conjugation<sup>45</sup>.

### 312 **Plasmid construction**

313 The dCas9-sgRNA plasmid expression system, pFR56, was derived from plasmid pJF1, a gift from Eligo  
314 Bioscience, harbouring a constitutively expressed sgRNA and cas9 under the control of a DAPG-inducible  
315 PhIF promoter<sup>41</sup>. This plasmid was recoded to avoid restriction sites and ensure plasmid stability in a  
316 maximum of *E. coli* strains<sup>42</sup>. We further modified this plasmid to inactivate Cas9 into dCas9 and to add the  
317 RP4 origin of transfer. Novel sgRNAs can be cloned on pFR56 using Golden Gate assembly<sup>58</sup> with BsaI  
318 restriction sites. The expression level of dCas9 on pFR56 was optimized to avoid the previously reported  
319 toxicity effect known as the bad-seed effect<sup>33</sup>. Briefly, we used the RBS calculator<sup>59</sup> to randomize 4  
320 positions of the dCas9 RBS and cloned the resulting library on the plasmid harboring an sgRNA with a bad-  
321 seed sequence (5'-TTGTATCAAACCATCACCCA-3') using the Gibson assembly method<sup>60</sup>. Candidate clones  
322 that grew normally in the presence of dCas9 induction were selected. In order to select clones that retain a  
323 sufficient dCas9 expression level for efficient repression of target genes, a sgRNA targeting the essential  
324 gene *rpsL* was cloned onto the psgRNACos vector (Addgene accession 114005) and introduced in the  
325 selected candidates. We discarded clones that were not killed in the presence of dCas9 induction. Finally  
326 the sgRNA was replaced by a *ccdB* counter-selection cassette in between two BsaI restriction sites<sup>61</sup>. This  
327 ensures the selection of clones in which a sgRNA was successfully added to the plasmid during library  
328 cloning. The sequence of pFR56 with a control non-targeting sgRNA was deposited on Genbank (accession  
329 MT412099).

330 The dUTPase from TA447, the TA system from E101 and the retron system from H120 were cloned using a  
331 three-fragment Gibson assembly<sup>60</sup> from (i) the pZS24-MSC1<sup>62</sup> plasmid, (ii) the LC-E75 strain (Addgene  
332 accession 115925) (iii) the genome of TA447, E101 or H120 respectively, using primers listed in  
333 **Supplementary Table 7**. The resulting vectors were named pFR67, pFR71 and pFR75 respectively, and  
334 comprise a pTet promoter, a Kanamycin resistance cassette and a pSC101 origin. As a control, we also  
335 cloned a GFP on the same backbone, yielding pFR66. The dUTPase from TA447 and the GFP were expressed

336 from the pTet promoter while the TA system from E101 and the retron system from H120 were expressed  
337 from their natural promoter.

### 338 **Library design**

339 We retrieved all *E. coli* complete genomes from GenBank Refseq (available in February 2018). We  
340 estimated genome similarity calculating the pairwise Mash distance (M) between all genomes using Mash  
341 v.2.0<sup>63</sup>. Importantly, the correlation between the Mash distance (M) and Average Nucleotide Identity (ANI)  
342 in the range of 90-100% has been shown to be very strong, with  $M \approx 1-ANI$  (ref<sup>63</sup>). All the resulting Mash  
343 distances between *E. coli* genomes are well below 0.05, in agreement with the assumption that they all  
344 belong to the same species. We removed 67 genomes that were too similar (MASH distance < 0.0001),  
345 mainly corresponding to different versions of K-12 and O157:H7. In this case, we picked the one present for  
346 a longer period of time in the databases. This resulted in a dataset of 370 completely assembled genomes  
347 for comparison<sup>18</sup>. Pan-genomes are the full complement of genes in the species (or dataset, or phylogroup)  
348 and were built by clustering homologous proteins into families. We determined the lists of putative  
349 homologs between pairs of genomes with MMseqs2 v.3.0<sup>64</sup> by keeping only hits with at least 80% identity  
350 and an alignment covering at least 80% of both proteins. Homologous proteins were then clustered by  
351 single-linkage<sup>65</sup>. From the resulting pangenome, we selected 3380 proteins present in more than 333/370  
352 genomes (90%) in up to 4 copies per genome.

353 For each gene and strain, all possible sgRNAs were listed by selecting the 20 NGG-proximal nucleotides on  
354 the coding strand. In order to avoid sgRNAs targeting regions with single-nucleotide variants, a first pre-  
355 selection step was performed for each gene in order to select up to 12 sgRNAs based on the number of  
356 targeted strains: (i) sgRNAs targeting the highest number of strains ( $N_{max}$ ) were first selected; (ii) if less than  
357 12 guides were obtained, sgRNAs targeting  $N_{max}-1$  strains were selected, then  $N_{max}-2$  strains, etc until 90% x  
358  $N_{max}$  strains; (iii) if less than 3 sgRNAs were selected after this process (possibly due to high rates of  
359 variants), the 3 sgRNAs with the highest number of targeted strains were selected; (iv) in order to select  
360 sgRNAs targeting the strains that may have been missed, we then selected the strains targeted by less than  
361 3 sgRNAs and performed a similar selection procedure: sgRNAs targeting the maximum number of missed  
362 strains ( $N_{max\_missed}$ ) were selected, followed by sgRNAs targeting  $N_{max\_missed}-1$  strains, etc, until 80% x  
363  $N_{max\_missed}$  strains. Finally, sgRNAs targeting less than 30 strains (~8%) were discarded.

364 After the preselection process, a penalty score was calculated from each sgRNA in order to select the best 3  
365 sgRNAs targeting each gene. This score takes into account, (i) off-target effects, (ii) predicted efficiency, (iii)  
366 number of targeted strains.

- 367 (i) For each sgRNA, we calculated the fraction of strains having another 11-nt match on the coding  
368 strand of a gene and the fraction of strains having a 9-nt match on any strand in a promoter  
369 (loosely defined as 100 nt before gene start). The 1<sup>st</sup> score was calculated as the sum of these  
370 fractions.
- 371 (ii) We used a recent model<sup>36</sup> which predicts the repression efficiency of sgRNAs based on fitness  
372 data obtained in a previous CRISPRi screen<sup>33</sup>. For each gene, the predicted sgRNA activity was  
373 normalized from 0 (highest activity) to 1 (lowest activity) and was then used as a 2<sup>nd</sup> score.
- 374 (iii) The number of targeted strains (with a full-length match) was reported for each sgRNA. For  
375 each gene, this number was normalized from 0 (sgRNA targeting the most strains) to 1 (virtually  
376 no strain targeted) and was then used as a 3<sup>rd</sup> score.

377 For each gene, all preselected sgRNAs were attributed a global penalty score by summing the 3 scores  
378 described above. A strong penalty was applied to guides carrying a 5-nt seed sequence among the 10  
379 strongest bad-seed sequences identified by Cui *et al.* (2018)<sup>33</sup> (AGGAA, TAGGA, ACCCA, TTGGA, TATAG,

380 GAGGC, AAAGG, GGGAT, TAGAC, GTCCT), so that they were only selected as a last resort. For each gene,  
381 sgRNAs were ranked by increasing global penalty score and the 3 best sgRNAs were selected (if available). If  
382 one of these 3 sgRNAs targeted less than 350 strains (95%), a 4<sup>th</sup> sgRNA was added. This process resulted in  
383 a library of 11,188 sgRNAs targeting conserved protein-coding genes.

384 We also designed sgRNAs targeting rRNAs, tRNAs and ncRNAs. Since rRNAs are highly conserved between  
385 all strains, it is very simple to select sgRNAs targeting all strains. However, it is complicated to assess their  
386 potential off-target activity due to their presence in many copies. We therefore selected all 109 sgRNAs  
387 targeting all strains. Similarly, homologous tRNAs have very similar nucleotide sequences, which makes it  
388 difficult to assess the off-target activity of each sgRNA. We therefore selected 131 sgRNAs targeting > 90%  
389 of strains (> 333 strains). ncRNAs are very diverse and their annotation can substantially differ between  
390 genomes. All annotated ncRNAs in all genomes were first listed. Then, ncRNAs which sequence is present in  
391 > 100 genomes were kept. For each of these ncRNAs, all possible sgRNAs were listed and assessed for off-  
392 target activity. For each ncRNA, we determined the smallest off-target size *s* for which at least 3 guides had  
393 no off-target activity (or in < 5% of strains). We then selected all the guides having an off-target of size *s* in <  
394 5% of strains.

395 Finally, we generated 20 non-targeting control sgRNAs. These guides should not have any on-target nor off-  
396 target activity in any strain. To ensure a minimal activity, we first generated all possible random 8-mers and  
397 kept those which never occur next to a PAM in a subset of 20 strains. We then generated 20 sgRNAs whose  
398 8 last 3' bases were randomly chosen from this subset, and whose 12 first 5' bases were random. We  
399 verified that each control sgRNA does not have an off-target in more than 1% of strains.

## 400 **Library construction**

401 The resulting library of 11,629 sgRNAs was generated as single-stranded DNA through on-chip oligo  
402 synthesis (CustomArray). Pooled oligo extension was performed with KAPA HiFi DNA polymerase (Roche)  
403 with primer FR222. The library was then amplified by PCR (KAPA HiFi polymerase; 95°C - 3'; 6 cycles 98°C -  
404 20", 60°C - 15", 72°C - 20"; 72°C - 10') with primers FR221 and FR222 and purified by gel extraction. The  
405 pFR56-ccdB vector was digested with BsaI (New England Biolabs) and gel purified. The plasmid library was  
406 then assembled using the Gibson method<sup>60</sup>.

407 During transformation, the initial absence of repressor proteins in the cell can result in a transient dCas9  
408 expression which can introduce a bias in the library. To avoid this, we built a library cloning strain, FR-E03,  
409 by integrating a constitutively-expressed PhIF repressor gene in the chromosome of MG1655. Briefly, a phIF  
410 expression cassette was cloned onto the pOSIP backbone<sup>66</sup> and integrated into HK022 *attP* site. The pOSIP  
411 backbone was then excised using the pE-FLP plasmid which was cured by serial restreaks. For library  
412 transformation, FR-E03 cells were grown in LB (200 mL) to OD ~ 1, washed 3 times in ice-cold pure water  
413 and resuspended in 250 µL ice-cold water. Ten electroporations were performed with 20 µL of cells and 0.5  
414 µL of dialyzed Gibson assembly product and pooled together. After 1h at 37°C, cells were plated on 10 large  
415 LB-Cm plates (12x12 cm) and incubated overnight at room temperature. The next day, each plate was  
416 washed twice with 5 mL LB-Cm and pooled. Plasmids were extracted by miniprep (Mancherey-Nagel)  
417 before further transformation into the conjugation strain MFDpir. This *pir+* strain is auxotrophic to  
418 diaminopimelic acid (DAP) and contains the RP4 conjugation machinery<sup>45</sup>. We attempted to integrate the  
419 same construction as in FR-E03 in the conjugation strain MFDpir but this was unsuccessful. Instead, we  
420 used pFR58, a low-copy pSC101 Kan<sup>R</sup> plasmid with the same PhIF expression system. We confirmed that  
421 pFR58 cannot be mobilized during conjugation since it does not contain the RP4 transfer machinery. The  
422 library was electroporated into MFDpir+pFR58 as described above. Transformants were selected on LB agar  
423 supplemented with Cm and 300 µM DAP and pooled before conjugation.

## 424 **Strain selection**

425 Starting from a collection of 92 *E. coli* natural isolates encompassing the phylogenetic diversity of the  
426 species and originating from various habitats in diverse conditions (environment, birds, non-human  
427 mammals and humans; gut commensalism, intestinal and extra-intestinal infections) (Clermont and  
428 Denamur, personal data), we performed growth curves to identify natural resistance to chloramphenicol  
429 which is the selection marker used in pFR56. Briefly, overnight cultures were diluted 100-fold in LB or in LB-  
430 Cm and OD600 was measured every 10 min for 8 h at 37°C with shaking on a Tecan Infinite M200Pro.  
431 Successful growth in Cm was observed in seven strains which were discarded. dCas9-mediated repression  
432 was then tested by conjugating pFR56 bearing an sgRNA targeting the essential gene *rpsL* into each strain.  
433 Plating on LB + Cm + 50 µM DAPG induced strong killing in all strains, suggesting that dCas9-mediated  
434 repression is functional in all strains. From the remaining isolates, we selected a panel of 18 strains  
435 including K-12 MG1655 from diverse origin and pathogenicity spanning most common *E. coli* phylogroups  
436 (A, B1, B2, D, E and F). Phylogroups were verified by quadruplex PCR with the Clermont method<sup>67</sup>. Strains  
437 selected for screening are listed in **Supplementary Table 1**.

## 438 **Bacterial conjugation**

439 MFDpir cells carrying the plasmid library were grown to OD~1 in LB-Cm supplemented with 300 µM DAP.  
440 Cells were then washed (2,000g – 10') to remove traces of Cm. Recipient strains were grown to stationary  
441 phase. Donor and recipients cells were mixed 1:1 (v/v, 0.1 to 1mL), pelleted (2,000g – 10 min), resuspended  
442 in 10-100 µL LB + 300 µM DAP, pipetted onto a LB + 300 µM DAP plate and incubated at 37°C for 2 h. As a  
443 negative control, donor and recipient strains were plated on a LB-Cm plate. For conjugation of individual  
444 sgRNAs, cells were then restreaked on LB-Cm to select individual transconjugants. For conjugation of the  
445 EcoCG library, cells were collected, resuspended in 1 mL of LB-Cm and plated on a large LB-Cm plate (12 x  
446 12 cm) followed by overnight incubation at room temperature. Ten-fold dilutions were also plated for CFU  
447 counting. We obtained >10<sup>7</sup> clones for each of the 18 strains assayed in this study, ensuring a > 1000-fold  
448 coverage. After overnight incubation at room temperature, plates containing nascent colonies were  
449 washed twice in 5 mL LB-Cm and stored at -80°C with 7.5% DMSO.

## 450 **Screen design**

451 Strains conjugated with the library were arrayed by mixing 150 µL of the -80°C stock with 1350 µL LB-Cm in  
452 duplicates on a 96-deepwell plate (Masterblock 96 well, 2ml, V-bottom plates by Greiner Bio-one). The  
453 plate was incubated overnight at 37°C in a Thermomixer (Eppendorf) with shaking (700 rpm). The next day,  
454 cultures were washed 1:1 in M9 medium before inoculation of 15 µL into 1485 µL of either LB, M9-glucose  
455 or GMM, supplemented with 50 µM DAPG without antibiotic selection. The remaining cultures were  
456 harvested and plasmids were extracted by miniprep to obtain reference samples for each strain. All screens  
457 were then performed in a Thermomixer at 37°C with shaking (700 rpm). Screens in LB and M9-glucose were  
458 performed in aerobic condition while screens in GMM were performed in an anaerobic chamber (80% N<sub>2</sub>,  
459 10% CO<sub>2</sub> and 10% H<sub>2</sub>). In all three cases, 3 passages were performed by diluting 15 µL of cells into 1485 µL  
460 of DAPG-containing fresh medium (1:100 dilution) in the same conditions, every 3.5 h for LB and every 12 h  
461 for M9-glucose and GMM. This represents a total of ~20 generations ( $\log_2(100^3) \approx 19.9$ ). Plasmids were  
462 finally extracted with a 96-well miniprep kit (Macherey-Nagel) to obtain the final distribution of the library  
463 for each strain and medium. Four strains were not assayed in M9 medium because of insufficient growth as  
464 previously described<sup>68</sup>.

## 465 **Illumina sample preparation and sequencing**

466 Library sequencing was performed as previously described<sup>33,34</sup>. Briefly, primers listed in **Supplementary**  
467 **Table 7** were used to perform two consecutive PCR reactions with KAPA HiFi polymerase (Roche). Starting  
468 from 100 ng of library plasmid, the first PCR (95°C – 3 min; 9 cycles [98°C – 20 s; 60°C – 15 s; 72°C – 20 s];



469 72°C – 10 min) is performed in a 30-μL reaction with 8.6 pmol of each primer. For the second PCR, a 20-μL  
470 mix containing 100 pmol of primers is added to the first PCR and the resulting 50-μL reaction is incubated  
471 (95°C – 3 min ; 9 cycles [98°C – 20 s; 66°C – 15 s; 72°C – 20 s]; 72°C – 10 min) to add the 2nd index and the  
472 flow cell attachment sequences. The resulting 354 bp-PCR DNA fragments were gel extracted. Samples  
473 were pooled (150 ng of each reference samples and 100 ng of other samples) and the final library  
474 concentration was determined by qPCR (KAPA Library Quantification Kit, Roche). Sequencing was  
475 performed on a NextSeq 500 benchtop sequencer (Illumina) using a custom protocol as previously  
476 described<sup>34</sup>. We obtained an average of 3 million reads per sample, representing an average coverage of ~  
477 260X.

## 478 **Data analysis**

479 Index sequences were used to de-multiplex the data into individual samples with a custom Python script.  
480 Reproducibility between experimental duplicates was very high (median Pearson's  $r = 0.988$ ) except for a  
481 replicate of strain ROAR 8 in LB that had very low read counts. This sample was discarded, while biological  
482 replicates from other strains were pooled into a single sample per strain for subsequent analyses. sgRNAs  
483 with less than 20 reads in the initial time point of a given strain were discarded in the corresponding strain  
484 (1.9% of the library on average). Samples were then normalized by sample size. The log<sub>2</sub>FC value was  
485 calculated for each guide  $g$  and strain  $s$  as follows ( $s_{initial}$  and  $s_{final}$  represent the normalized reads  
486 counts of strain  $s$  in the initial and final time point respectively):

$$\log_2FC_{g,s} = \log_2 \left( \frac{Reads_{g,s_{final}} + 1}{Reads_{g,s_{initial}} + 1} \right)$$

487 The limit in sequencing depth imposes that sgRNAs having 20 initial reads and inducing a major fitness  
488 defect may be eliminated from the population after 20 generations. In this case, the limit in log<sub>2</sub>FC can be  
489 calculated as  $\log_2(1/21) = -4.4$ , which is sufficient to classify the targeted gene as essential using our  
490 threshold. We mapped the EcoCG library to each genome to identify sgRNAs which do not have a full-  
491 length match (e.g. because of single-nucleotide variants), and their log<sub>2</sub>FC value was set to NaN in the  
492 corresponding strain in order to avoid false negatives. Finally, the median log<sub>2</sub>FC were centered on the  
493 median log<sub>2</sub>FC of 20 control non-targeting sgRNAs, and the resulting values were used as gene scores. We  
494 selected genes as “variably essential” in **Supplementary Fig. 5** when repression induced a fitness defect in  
495 at least one strain (gene score < -3) and no fitness defect in at least one strain (gene score > -1). For the  
496 heatmaps drawn in **Fig. 4**, we used a more stringent threshold: for each gene, we calculated the minimum  
497 and maximum gene scores across strains after excluding those that had more than 50% of sgRNAs with  
498 missing values. We then kept genes whose minimal gene score was lower than -5 and whose maximal gene  
499 score was greater than -1 across all strains. Finally, in **Supplementary Fig. 7b**, we aimed at analyzing the  
500 relationship between phylogeny and essentiality between pairs of strains. Therefore, we selected “variably  
501 essential” genes that were essential (gene score < -3) in at least 2 strains and nonessential (gene score > -1)  
502 in at least 2 strains. From this subset, we also discarded genes when their effect could clearly be attributed  
503 to a polar effect on a downstream essential gene (for instance we discarded *ycaR* whose effect is due to a  
504 polar effect on *kdsB*). When analyzing the data, it is indeed important to consider that dCas9 repression of  
505 a gene in an operon will also silence all downstream genes. The synteny of core genes is strongly conserved  
506 between strains since most variations in gene content occur in hotspots<sup>17</sup>. Therefore, we expect most polar  
507 effects to be conserved between all strains. Polar effects should thus not be the source of differences in  
508 essentiality between strains in our dataset.

## 509 **Comparative genomics**

510 The genomes of the 18 strains used for screening were reannotated with Prokka 1.14.2 using default  
511 settings<sup>69</sup>. Proteins from these 18 strains were clustered using MMseqs2 v.3.0 with default parameters<sup>64</sup>.  
512 The resulting clusters were used to generate the core and pan-genome shown in **Fig. 2** using up to 250  
513 permutations of sets of strains (**Supplementary Fig. 7c-d**). To obtain pairwise phylogenetic distances  
514 between strains, we generated a core genome alignment with Parsnp<sup>70</sup> which was used to build a  
515 phylogenetic tree with FastTree2<sup>71</sup>. To evaluate the presence of homologs of essential genes, we clustered  
516 proteins from all 18 strains with a 40%-identity threshold with MMseqs2 v.3.0<sup>64</sup> (--min-seq-id 0.4) to obtain  
517 groups of sequence homologs. We then selected clusters containing essential proteins from K12-BW25113  
518 in LB<sup>13</sup>. To do so, we clustered proteins from BW25113 and MG1655 (--min-seq-id 0.95) to obtain a  
519 correspondence table between the names of BW25113 essential genes and MG1655 locus tags. We finally  
520 selected protein clusters containing at least one sequence per strain with at least one strain having more  
521 than one sequence.

522 We used sequence searches on the web interface of InterPro<sup>72</sup> and pfam<sup>73</sup> databases to look for known  
523 domains in protein candidates. Structural predictions were performed with Phyre2<sup>74</sup>. Phaster<sup>75</sup> was used  
524 with default parameters to identify prophages in the genome of strains E101 and H120.

## 525 **Screen results validation**

526 **sgRNA cloning.** Individual sgRNAs listed in **Supplementary Table 8** were cloned into pFR56 using Golden  
527 Gate assembly<sup>58</sup>. All constructions were validated by Sanger sequencing. Cloning was performed in MG1655  
528 or MFDpir before transfer to the appropriate strains by conjugation.

529 **Gene deletions.** *dut* was deleted from strain TA447 using the  $\lambda$ -red recombination system as described  
530 previously<sup>34</sup> using primers listed in **Supplementary Table 7**. We performed whole-genome sequencing  
531 (WGS) to verify the absence of compensatory mutations. We used breseq (v. 0.33.2) for variant calling<sup>76</sup>.

532 **Growth curves.** An overnight culture was washed in the appropriate medium to avoid nutrient carryover  
533 and diluted 1000-fold. Growth was monitored in triplicates by measuring optical density at 600 nm on an  
534 Infinite M200Pro (Tecan) at 37°C with shaking.

535 **RT-qPCR.** Overnight cultures of S88 carrying pFR56 or pFR56.27 (i.e. pFR56 with *ybaQ* sgRNA) in LB + Cm  
536 were diluted 1000-fold in 2 mL of LB + Cm + 50  $\mu$ M DAPG. An overnight culture of TA447 was diluted 1000-  
537 fold in 2 mL of LB and 100-fold in 2 mL of GMM to account for the slower growth rate in GMM. After 3h at  
538 37°C, RNAs were extracted using Trizol. RNA samples were treated with DNase (Roche) and reverse-  
539 transcribed into cDNA using the Transcriptor First Strand cDNA Synthesis Kit (Roche). qPCR was performed  
540 in two technical replicates with the FastStart Essential DNA Green master mix (Roche) on a LightCycler 96  
541 (Roche). Relative gene expression was computed using the  $\Delta\Delta$ Cq method after normalization by 5S rRNA  
542 (*rrsA*). qPCR primers are listed in **Supplementary Table 9**.

543 **Isolation of suppressor mutants.** To isolate suppressor mutants, we conjugated pFR56 harboring the  
544 corresponding guide into the appropriate strain and we selected clones that grew robustly with 50  $\mu$ M  
545 DAPG. In order to avoid selecting mutations on the plasmid that inactivate the CRISPRi system, we  
546 conjugated a second plasmid (pFR59) identical to pFR56, but carrying a kanamycin resistance cassette  
547 instead of the chloramphenicol resistance cassette. The resistance to repression was verified by plating  
548 serial dilutions of the transconjugants on LB agar plates with Kan  $\pm$  50  $\mu$ M DAPG. In the case of JJ1886, this  
549 strain is naturally resistant to kanamycin. We therefore built a third plasmid (pFR72) with a gentamycin  
550 resistance cassette and used it for the first selection step together with pFR56 in the second selection step.  
551 Finally, to avoid selecting clones that acquired mutations in the chromosomal sgRNA target, we performed

552 Sanger sequencing on the genomic region flanking the sgRNA binding site and discarded clones with  
553 mutations in the target. Genomic DNA was extracted from selected clones as well as in the parental clone  
554 using the Wizard Genomic DNA Purification Kit (Promega). NGS was performed using Nextera XT DNA  
555 Library Preparation kit and the NextSeq 500 sequencing systems (Illumina) at the Mutualized Platform for  
556 Microbiology (P2M) at Institut Pasteur. Mutations were identified by mapping raw reads to the appropriate  
557 genome using breseq v. 0.33.2<sup>76</sup>. Among the genomes we sequenced, APEC O1 had a previously unreported  
558 plasmid. The new genome was deposited on the European Nucleotide Archive (ENA) under the accession  
559 GCA\_902880315. We also found that the previously reported genome sequence of H120  
560 (GCF\_000190855.1) had a high number of sequencing errors introducing frameshifts and premature stop  
561 codons. We re-sequenced this strain to correct these errors and deposited the resulting corrected genome  
562 sequence on the ENA with the accession GCA\_902876715. We also resequenced our clone of K-12 MG1655  
563 and deposited the genome on the ENA with the accession ERS5065070.

564 **Bacteriophage efficiency of plaquing.** To test the effect of different systems on phage resistance, 250 µL of  
565 overnight cultures of MG1655 carrying either pFR66, pFR71 or pFR75 were mixed with CaCl<sub>2</sub> (5 mM final) in  
566 12 mL of top agar (LB + 0.5% agar). The mixture was poured onto a large square LB + Kan plate. Stocks of  
567 phages λ, T4, T7, P1vir, 186cIts were serially diluted in PBS and 2 µL of each dilution were spotted on the  
568 bacterial lawns. Plates were then incubated overnight at 37°C and plaque-forming units were counted the  
569 next day.

## 570 **RNA-seq**

571 Overnight cultures were diluted 100-fold in 1 mL of LB in a 96-deepwell plate (Masterblock 96 well, 2ml, V-  
572 bottom plates by Greiner Bio-one). After 2.5 h at 37°C, cultures were diluted to OD ~ 0.02 in 1.4 mL of LB  
573 and were further grown for 2 h at 37°C on a Thermomixer (Eppendorf) with shaking (700 rpm). Each culture  
574 was then transferred to a 2-mL Eppendorf containing 170 µL of stop solution (5% acid phenol in ethanol)  
575 and cooled down for 10 seconds in a bath of dry ice and ethanol. Cells were harvested by centrifugation at  
576 0°C (1 min – 16,000 g) and the pellets were frozen at -80°C. For RNA extraction, pellets were thawed on ice,  
577 resuspended in 200 µL of pre-warmed lysozyme solution and incubated for 3 min at 37°C before addition of  
578 1 mL of Trizol. Samples were vigorously vortexed and incubated at room temperature for 5 min followed by  
579 addition of 200 µL of chloroform. After vigorous vortexing, samples were incubated for 5 min at room  
580 temperature and centrifuged (10 min – 12,000 g) to separate phases. The upper aqueous phase was  
581 collected and RNA was precipitated by addition of 500 µL of isopropanol. Samples were incubated for 10  
582 min at room temperature before centrifugation (10 min – 12,000 g). Pellets were washed with 1 mL of 75%  
583 ethanol and centrifuged (5 min – 7,500 g). The pellets were finally air-dried and resuspended in 50 µL of  
584 pure water. RNA samples were DNase-treated using TURBO DNA-free Kit (Thermo Fisher Scientific) and  
585 sample quality was assessed on an Agilent Bioanalyzer 2100. Samples were prepared for sequencing using  
586 the TruSeq® Stranded Total RNA Kit (Illumina) and sequenced on a NextSeq 500 benchtop sequencer  
587 (Illumina). Raw reads were aligned on each genome using Bowtie2 v2.3.4.3<sup>77</sup>. Alignment files were  
588 converted with Samtools v1.9<sup>78</sup> and read counts for each gene were obtained using HTseq v0.9.1<sup>79</sup>. Raw  
589 read counts were normalized by sample size and by gene length to obtain reads-per-kilobase-per-million  
590 (RPKM). The log<sub>2</sub>-transformed median RPKM value of each strain across biological replicates was used as a  
591 measure of gene expression in each strain.

592

## 593 **Code availability**

594 Custom scripts used in the manuscript can be found here: <https://gitlab.pasteur.fr/dbikard/ecocg>.

## 595 **Data availability**

596 Raw sequencing reads from CRISPRi screens are available at the European Nucleotide Archive under the  
597 accession PRJEB37847. Raw reads from RNA-seq experiments were deposited on ArrayExpress with the  
598 accession E-MTAB-9036. Processed data is available in Supplementary Tables.

## 599 **References**

- 600 1. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat. Rev.*  
601 *Genet.* **19**, 34–49 (2017).
- 602 2. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than  
603 are nonessential genes in bacteria. *Genome Res.* **12**, 962–8 (2002).
- 604 3. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420  
605 (2015).
- 606 4. Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L. & Whiteley, M. Essential genome of *Pseudomonas*  
607 *aeruginosa* in cystic fibrosis sputum. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4110–4115 (2015).
- 608 5. Le Breton, Y. *et al.* Essential Genes in the Core Genome of the Human Pathogen *Streptococcus pyogenes*. *Sci.*  
609 *Rep.* **5**, 9838 (2015).
- 610 6. Freed, N. E., Bumann, D. & Silander, O. K. Combining *Shigella* Tn-seq data with gold-standard *E. coli* gene  
611 deletion data suggests rare transitions between essential and non-essential gene functionality. *BMC Microbiol.*  
612 **16**, 203 (2016).
- 613 7. Poulsen, B. E. *et al.* Defining the core essential genome of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S.*  
614 *A.* **116**, 10072–10080 (2019).
- 615 8. Galardini, M. *et al.* The impact of the genetic background on gene deletion phenotypes in *Saccharomyces*  
616 *cerevisiae*. *Mol. Syst. Biol.* **15**, (2019).
- 617 9. Dowell, R. D. *et al.* Genotype to phenotype: a complex problem. *Science* **328**, 469 (2010).
- 618 10. van Opijnen, T., Dedrick, S. & Bento, J. Strain Dependent Genetic Networks for Antibiotic-Sensitivity in a  
619 Bacterial Pathogen with a Large Pan-Genome. *PLoS Pathog.* **12**, e1005869 (2016).
- 620 11. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio  
621 collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
- 622 12. Nichols, R. J. *et al.* Phenotypic Landscape of a Bacterial Cell. *Cell* **144**, 143–156 (2011).
- 623 13. Goodall, E. C. A. *et al.* The Essential Genome of *Escherichia coli* K-12. *MBio* **9**, e02096-17 (2018).
- 624 14. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–  
625 509 (2018).
- 626 15. Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-  
627 coded transposons. *MBio* **6**, e00306-15 (2015).
- 628 16. Rasko, D. A. *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli*  
629 commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–93 (2008).
- 630 17. Touchon, M. *et al.* Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse  
631 Adaptive Paths. *PLoS Genet.* **5**, e1000344 (2009).
- 632 18. Touchon, M. *et al.* Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*.  
633 *PLoS Genet.* **16**, e1008866 (2020).
- 634 19. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli*. *Nat.*  
635 *Rev. Microbiol.* **8**, 207–217 (2010).

- 636 20. Denamur, E., Clermont, O., Bonacorsi, S. & Gordon, D. The population genetics of pathogenic *Escherichia coli*.  
637 *Nat. Rev. Microbiol.* 1–18 (2020). doi:10.1038/s41579-020-0416-x
- 638 21. Subashchandrabose, S., Smith, S. N., Spurbeck, R. R., Kole, M. M. & Mobley, H. L. T. Genome-Wide Detection  
639 of Fitness Genes in Uropathogenic *Escherichia coli* during Systemic Infection. *PLOS Pathog.* **9**, e1003788  
640 (2013).
- 641 22. Olson, M. A., Siebach, T. W., Griffiths, J. S., Wilson, E. & Erickson, D. L. Genome-Wide Identification of Fitness  
642 Factors in Mastitis-Associated *Escherichia coli*. *Appl. Environ. Microbiol.* **84**, e02190-17 (2018).
- 643 23. Phan, M.-D. *et al.* The Serum Resistome of a Globally Disseminated Multidrug Resistant Uropathogenic  
644 *Escherichia coli* Clone. *PLOS Genet.* **9**, e1003834 (2013).
- 645 24. Goh, K. G. K. *et al.* Genome-Wide Discovery of Genes Required for Capsule Production by Uropathogenic  
646 *Escherichia coli*. *MBio* **8**, e01558-17 (2017).
- 647 25. Shea, A. E. *et al.* Identification of *Escherichia coli* CFT073 fitness factors during urinary tract infection using an  
648 ordered transposon library. *Appl. Environ. Microbiol.* (2020). doi:10.1128/AEM.00691-20
- 649 26. Bergmiller, T., Ackermann, M. & Silander, O. K. Patterns of Evolutionary Conservation of Essential Genes  
650 Correlate with Their Compensability. *PLoS Genet.* **8**, e1002803 (2012).
- 651 27. Patrick, W. M., Quandt, E. M., Swartzlander, D. B. & Matsumura, I. Multicopy Suppression Underpins  
652 Metabolic Evolvability. *Mol. Biol. Evol.* **24**, 2716–2722 (2007).
- 653 28. Martínez-Carranza, E. *et al.* Variability of Bacterial Essential Genes Among Closely Related Bacteria: The Case  
654 of *Escherichia coli*. *Front. Microbiol.* **9**, 1059 (2018).
- 655 29. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene  
656 Expression. *Cell* **152**, 1173–1183 (2013).
- 657 30. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered  
658 CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).
- 659 31. Vigouroux, A. & Bikard, D. CRISPR Tools To Control Gene Expression in Bacteria. *Microbiol. Mol. Biol. Rev.* **84**,  
660 (2020).
- 661 32. Rousset, F. & Bikard, D. CRISPR screens in the era of microbiomes. *Curr. Opin. Microbiol.* **57**, 70–77 (2020).
- 662 33. Cui, L. *et al.* A CRISPRi screen in *E. coli* reveals sequence-specific toxicity of dCas9. *Nat. Commun.* **9**, 1912  
663 (2018).
- 664 34. Rousset, F. *et al.* Genome-wide CRISPR-dCas9 screens in *E. coli* identify essential genes and phage host factors.  
665 *PLOS Genet.* **14**, e1007749 (2018).
- 666 35. Wang, T. *et al.* Pooled CRISPR interference screening enables genome-scale functional genomics study in  
667 bacteria with superior performance. *Nat. Commun.* **9**, 2475 (2018).
- 668 36. Calvo-Villamañán, A. *et al.* On-target Activity Predictions Enable Improved CRISPR-dCas9 Screens in Bacteria.  
669 *Nucleic Acids Res.* **48**, e64 (2020).
- 670 37. Li, S. *et al.* Genome-Wide CRISPRi-Based Identification of Targets for Decoupling Growth from Production. *ACS*  
671 *Synth. Biol.* **9**, 1030–1040 (2020).
- 672 38. Lee, H. H. *et al.* Functional genomics of the rapidly replicating bacterium *Vibrio natriegens* by CRISPRi. *Nat.*  
673 *Microbiol.* **4**, 1105–1113 (2019).
- 674 39. Yao, L. *et al.* Pooled CRISPRi screening of the cyanobacterium *Synechocystis* sp PCC 6803 for enhanced  
675 industrial phenotypes. *Nat. Commun.* **11**, 1666 (2020).
- 676 40. Liu, X., Kimmey, J. M., Bakker, V. de, Nizet, V. & Veening, J.-W. Exploration of bacterial bottlenecks and  
677 *Streptococcus pneumoniae* pathogenesis by CRISPRi-seq. *bioRxiv* 2020.04.22.055319 (2020).  
678 doi:10.1101/2020.04.22.055319

- 679 41. Schnider-Keel, U. *et al.* Autoinduction of 2,4-diacetylphloroglucinol biosynthesis in the biocontrol agent  
680 *Pseudomonas fluorescens* CHA0 and repression by the bacterial metabolites salicylate and pyoluteorin. *J.*  
681 *Bacteriol.* **182**, 1215–1225 (2000).
- 682 42. Decrulle, A., Fernandez Rodriguez, J., Duportet, X. & Bikard, D. OPTIMIZED VECTOR FOR DELIVERY IN  
683 MICROBIAL POPULATIONS. (2018).
- 684 43. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated  
685 Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837 (1988).
- 686 44. Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and  
687 manipulated in gnotobiotic mice. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6252–7 (2011).
- 688 45. Ferrières, L. *et al.* Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli*  
689 random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range  
690 RP4 conjugative machinery. *J. Bacteriol.* **192**, 6418–27 (2010).
- 691 46. Rocha, E. P. C. & Danchin, A. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial  
692 Proteins. *Mol. Biol. Evol.* **21**, 108–116 (2004).
- 693 47. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J.*  
694 *Mol. Biol.* **333**, 863–882 (2003).
- 695 48. Tye, B.-K. & Lehman, I. R. Excision repair of uracil incorporated in DNA as a result of a defect in dUTPase. *J.*  
696 *Mol. Biol.* **117**, 293–306 (1977).
- 697 49. Schaub, R. E. & Hayes, C. S. Deletion of the RluD pseudouridine synthase promotes SsrA peptide tagging of  
698 ribosomal protein S7. *Mol. Microbiol.* **79**, 331–341 (2011).
- 699 50. Luo, P., He, X., Liu, Q. & Hu, C. Developing Universal Genetic Tools for Rapid and Efficient Deletion Mutation in  
700 *Vibrio* Species Based on Suicide T-Vectors Carrying a Novel Counterselectable Marker, vmi480. *PLoS One* **10**,  
701 e0144465 (2015).
- 702 51. Aakre, C. D., Phung, T. N., Huang, D. & Laub, M. T. A bacterial toxin inhibits DNA replication elongation through  
703 a direct interaction with the  $\beta$  sliding clamp. *Mol. Cell* **52**, 617–628 (2013).
- 704 52. Harms, A., Brodersen, D. E., Mitarai, N. & Gerdes, K. Toxins, Targets, and Triggers: An Overview of Toxin-  
705 Antitoxin Biology. *Molecular Cell* **70**, 768–784 (2018).
- 706 53. Burroughs, A. M., Zhang, D., Schäffer, D. E., Iyer, L. M. & Aravind, L. Comparative genomic analyses reveal a  
707 vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids*  
708 *Res.* **43**, 10633–54 (2015).
- 709 54. Bobonis, J. *et al.* Bacterial retrons encode tripartite toxin/antitoxin systems. *bioRxiv* 2020.06.22.160168  
710 (2020). doi:10.1101/2020.06.22.160168
- 711 55. Bobonis, J. *et al.* Phage proteins block and trigger retron toxin/antitoxin systems. *bioRxiv* 2020.06.22.160242  
712 (2020). doi:10.1101/2020.06.22.160242
- 713 56. Millman, A. *et al.* Bacterial retrons function in anti-phage defense. *bioRxiv* 2020.06.21.156273 (2020).  
714 doi:10.1101/2020.06.21.156273
- 715 57. Gao, L. *et al.* Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* **369**, 1077–1084  
716 (2020).
- 717 58. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden Gate Shuffling: A One-Pot DNA Shuffling Method  
718 Based on Type IIs Restriction Enzymes. *PLoS One* **4**, e5553 (2009).
- 719 59. Salis, H. M. The ribosome binding site calculator. in *Methods in Enzymology* **498**, 19–42 (Academic Press Inc.,  
720 2011).
- 721 60. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**,  
722 343–345 (2009).

- 723 61. Hartley, J. L., Temple, G. F. & Brasch, M. A. DNA cloning using in vitro site-specific recombination. *Genome Res.* **10**, 1788–95 (2000).  
724
- 725 62. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in escherichia coli via the  
726 LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
- 727 63. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**,  
728 132 (2016).
- 729 64. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive  
730 data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
- 731 65. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 1–8 (2018).
- 732 66. St-Pierre, F. *et al.* One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013).
- 733 67. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont Escherichia coli phylo-typing  
734 method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **5**,  
735 58–65 (2013).
- 736 68. Bouvet, O., Bourdelier, E., Glodt, J., Clermont, O. & Denamur, E. Diversity of the auxotrophic requirements in  
737 natural isolates of Escherichia coli. *Microbiology* **163**, 891–899 (2017).
- 738 69. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
- 739 70. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment  
740 and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 524 (2014).
- 741 71. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large  
742 Alignments. *PLoS One* **5**, e9490 (2010).
- 743 72. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence  
744 annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
- 745 73. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
- 746 74. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein  
747 modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
- 748 75. Arndt, D. *et al.* PHASTER: a better, faster version of the PFAST phage search tool. *Nucleic Acids Res.* **44**, W16–  
749 W21 (2016).
- 750 76. Deatherage, D. E. & Barrick, J. E. Identification of Mutations in Laboratory-Evolved Microbes from Next-  
751 Generation Sequencing Data Using breseq. in *Engineering and analyzing multicellular systems* 165–188  
752 (Humana Press, New York, NY, 2014). doi:10.1007/978-1-4939-0554-6\_12
- 753 77. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
- 754 78. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- 755 79. Anders, S., Pyl, P. T. & Huber Wolfgang. HTSeq—a Python framework to work with high-throughput  
756 sequencing data. *Bioinformatics* **31**, 166–169 (2015).

757

## 758 **Acknowledgement**

759 We thank Bruno Dupuy for sharing the use of the anaerobic chamber, as well as Olivier Tenaillon, Pierre-Alexandre  
760 Kaminsky, Guennadi Sezonov, Antoine Danchin and Alicia Calvo-Villamañan for useful discussions. We thank the P2M  
761 platform (Institut Pasteur, Paris, France) for genome sequencing and Valérie Briolat from the Biomics platform, C2RT,  
762 Institut Pasteur, Paris, France, supported by France Génomique (ANR-10-INBS-09-09) and IBISA. We also thank  
763 Jerónimo Rodríguez-Beltrán and Sylvain Brisse for providing a strain of *Citrobacter freundii* and *Klebsiella pneumoniae*  
764 respectively. This work was supported by the European Research Council (ERC) under the Europe Union’s Horizon



765 2020 research and innovation program (grant agreement No [677823]), by the French Government's Investissement  
766 d'Avenir program and by Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (ANR-10-LABX-  
767 62-IBEID), F.R. is supported by a doctoral scholarship from Ecole Normale Supérieure. E.D. was partially supported by  
768 the "Fondation pour la Recherche Médicale" (Equipe FRM 2016, grant number DEQ20161136698). E.R. was partially  
769 supported by the "Fondation pour la Recherche Médicale" (Equipe FRM EQU201903007835).

#### 770 **Author contributions**

771 F.R. and D.B. designed the project. E.R. performed bioinformatic computation of the *E. coli* pangenome. E.D. and O.C.  
772 provided strains and genome sequences. F.R. performed experiments and analyzed data. J.R.F. and F.P.F. participated  
773 in the design of pFR56. J.C.C. provided experimental assistance. F.R., E.R. and D.B. wrote the manuscript. D.B.  
774 supervised the project.

#### 775 **Competing interests**

776 The authors declare no competing interests.