

1 **Genome-wide CRISPR-Cas9 screen in *E. coli* identifies design rules**  
2 **for efficient targeting**

3 Belen Gutierrez†, Jérôme Wong Ng†, Lun Cui, Christophe Becavin & David Bikard\*

4 Synthetic Biology Group, Microbiology Department, Institut Pasteur, Paris, 75015, France

5 \* To whom correspondence should be addressed. Tel: +33140613924; Email:

6 david.bikard@pasteur.fr

7 † Belen Gutierrez and Jérôme Wong Ng Contributed equally to this work.

8

9 **Running title:** Predicting CRISPR-Cas9 activity in *E. coli*

10 **Keywords:** CRISPR-Cas9, *Escherichia coli*, antimicrobial, genome-wide screen

## 11 **Abstract**

12 **The main outcome of efficient CRISPR-Cas9 cleavage in the chromosome of bacteria is cell death.**  
13 **This can be conveniently used to eliminate specific genotypes from a mixed population of bacteria,**  
14 **which can be achieved both *in vitro*, e.g. to select mutants, or *in vivo* as an antimicrobial strategy.**  
15 **The efficiency with which Cas9 kills bacteria has been observed to be quite variable depending on**  
16 **the specific target sequence, but little is known about the sequence determinants and mechanisms**  
17 **involved. Here we performed a genome-wide screen of Cas9 cleavage in the chromosome of *E. coli***  
18 **to determine the efficiency with which each guide RNA kills the cell. Surprisingly we observed a**  
19 **large-scale pattern where guides targeting some regions of the chromosome are more rapidly**  
20 **depleted than others. Unexpectedly, this pattern arises from the influence of degrading specific**  
21 **chromosomal regions on the copy number of the plasmid carrying the guide RNA library. After**  
22 **taking this effect into account, it is possible to train a neural network to predict Cas9 efficiency**  
23 **based on the target sequence. We show that our model learns different features than previous**  
24 **models trained on Eukaryotic CRISPR-Cas9 knockout libraries. Our results highlight the need for**  
25 **specific models to design efficient CRISPR-Cas9 tools in bacteria.**

## 26 **Introduction**

27 The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) -associated protein 9 (Cas9)  
28 has been repurposed as a tool for a variety of applications including genome editing, transcriptional  
29 repression or activation, epigenetic modifications, chromosomal loci tagging and more (Hsu, Lander,  
30 and Zhang 2014). In most Eukaryotic cells, Cas9 breaks can be efficiently repaired by template-  
31 independent non-homologous end joining (NHEJ), which introduces small indels and can be  
32 conveniently used to knockout genes. Conversely, the main outcome of Cas9 cleavage in the  
33 chromosome of bacteria is cell death (Bikard et al. 2012, 2014; Citorik, Mimee, and Lu 2014; Gomaa  
34 et al. 2013). Most bacteria lack a NHEJ system, but even in species that do carry such repair pathway,  
35 it seems unable to efficiently repair Cas9 break under laboratory conditions (T. Xu et al. 2015;  
36 Bernheim et al. 2017). Bacteria mostly rely on homologous recombination to repair double strand  
37 breaks, which is made possible by the fact that several copies of the chromosome are present in the  
38 cell under most conditions (Dillingham and Kowalczykowski 2008; Ayora et al. 2011).

39 We recently showed that when Cas9 is guided to target a position in the chromosome, some guide  
40 RNAs appear to be less efficient than others. When guided by a weak guide, Cas9 will not cleave all  
41 copies of the chromosome simultaneously leaving a copy intact for repair and allowing cells to  
42 survive by entering a cycle of DNA cleavage and repair. However, a strong guide will lead to the

43 simultaneous cleavage of all chromosome copies making repair through homologous recombination  
44 impossible and leading to cell death (Cui and Bikard 2016). This property can be used to select  
45 specific mutants or genotypes in mixed bacterial populations or even harnessed to engineer  
46 sequence-specific antimicrobials able to kill target antibiotic resistant or virulent bacteria (Bikard et  
47 al. 2014; Gomaa et al. 2013; Jiang et al. 2013; Citorik, Mimee, and Lu 2014). For all these applications  
48 it is critical to understand what makes some guide RNAs better at killing the cell than others. Cas9  
49 cleavage is the result of a process which starts with the formation of the gRNA-Cas9 nucleoprotein  
50 complex. Target search then occurs through the recognition of a small protospacer adjacent motif  
51 (PAM), followed by R-loop formation through pairing of the guide RNA with the target strand (S. H.  
52 Sternberg et al. 2014; Anders et al. 2014; Jinek et al. 2012). Upon successful pairing a conformational  
53 shift in the Cas9 protein occurs bringing two catalytic domains, RuvC and HNH, in contact with the  
54 DNA and leading to the formation of a double strand break (Samuel H. Sternberg et al. 2015; Jinek et  
55 al. 2014). Each of these steps could be impacted by the sequence of the guide RNA, genomic context,  
56 DNA conformation and DNA modifications.

57 High-throughput screens performed in Eukaryotic systems have enabled to identify sequence  
58 features that determine knockout efficiency (Doench et al. 2014, 2016; H. Xu et al. 2015; Moreno-  
59 Mateos et al. 2015; Wang et al. 2014). However, the features that emerge in these models seem to  
60 be very different between experimental setups, and likely result in great part from constraints on the  
61 proper expression and stability of guide RNAs, rather than from their impact on Cas9 biochemical  
62 activity (Haeussler et al. 2016; Moreno-Mateos et al. 2015).

63 The goal of this study is to elucidate the genetic requirements for efficient CRISPR-Cas9 targeting in  
64 the bacterial chromosome. To this end, we performed a high-throughput screen in which we guide  
65 Cas9 to cleave ~92,000 different random positions around the chromosome of *E. coli* MG1655,  
66 followed by a “NGG” PAM. Upon Cas9 induction, guides that efficiently kill *E. coli* are expected to be  
67 rapidly depleted from the library, which can be measured through sequencing of the guide RNA  
68 library before and after induction of Cas9 expression. We observed large regions along the  
69 chromosome (~100 kb) in which the sgRNAs are rapidly depleted from the library while sgRNAs from  
70 other regions, are still present several hours after Cas9 induction. Surprisingly, this large-scale  
71 pattern is not correlated to cell death, as revealed by time-lapse microscopy. Indeed, cells  
72 immediately stop dividing, start to filament and then die after Cas9 induction regardless of the region  
73 targeted. We reveal that the pattern observed is rather due to differences in the copy number of the  
74 plasmid carrying the guide RNA library that arise after DNA cleavage in different regions of the  
75 chromosome. The loss of some genomic regions leads to an interruption of plasmid replication, while  
76 the plasmid can still replicate and increase its copy number after cleavage in other regions. The scale

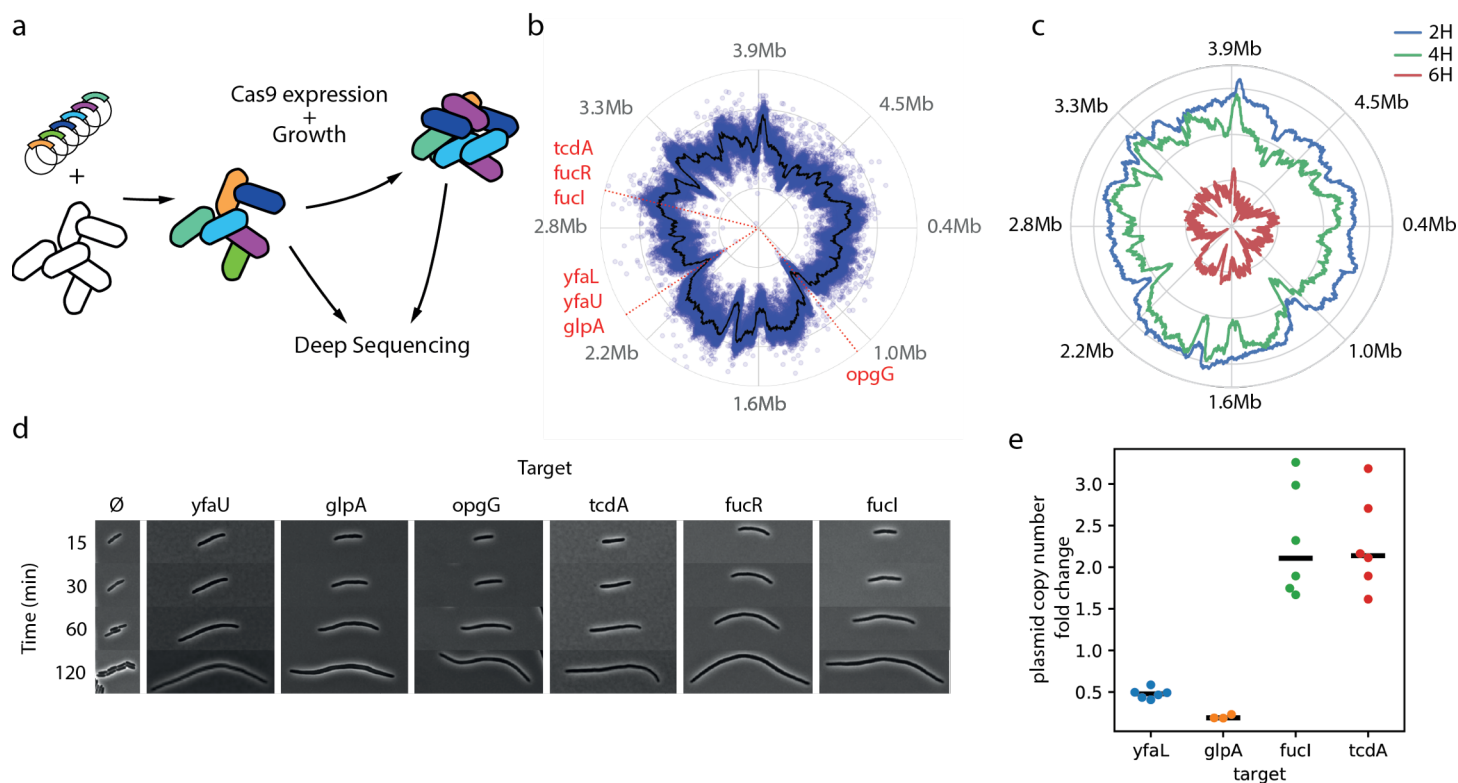
77 and shape of the pattern is determined by the extent of chromosomal DNA being degraded after  
78 Cas9 cleavage. After taking this interesting phenomenon into account, one can still observe  
79 differences in the efficiency of guide RNAs targeting the chromosome of *E. coli*. We built a neural  
80 network model able to predict this efficiency based on the guide sequence. Model previously build  
81 on Eukaryotic datasets have a limited predictive power on our data, highlighting the need to develop  
82 specific models of guide RNA activity in bacteria. Our model should directly aid in the design of guide  
83 RNA for genome editing and the development of CRISPR antimicrobials. We are making it available as  
84 guide RNA design tool at the following address: <http://hub13.hosting.pasteur.fr:8080/CRISPRBact/>.

## 85 **Results**

### 86 **Large-scale depletion pattern of sgRNAs after Cas9 cleavage**

87 We previously constructed a library containing 92,000 guides targeting random positions in the  
88 genome of *E. coli* followed by a “NGG” PAM motif (Cui et al. 2018). We introduced this library in *E.*  
89 *coli* MG1655 carrying Cas9 in the chromosome under the control of a Ptet promoter (strain LC-E19).  
90 Cas9 was induced and the library sequenced before induction and at different time points in order to  
91 determine the efficiency of killing of each guide (**Fig. 1a**). The experiment was performed in  
92 triplicates and the depletion of guides in the library was computed as the log<sub>2</sub> transformed fold  
93 change of read counts normalized to a non-targeting guide RNA. When plotting the results along the  
94 chromosome of *E. coli*, we immediately noticed a large-scale pattern where guides targeting certain  
95 regions on a scale of ~100 Kbp are depleted from the population faster than guides targeting other  
96 regions. This pattern can be represented as the moving average of log<sub>2</sub>FC with a sliding window of  
97 6kb, and is especially striking after 4H of induction (**Fig. 1b, c**).

98 We initially formulated the hypothesis that some chromosomal regions might be less accessible to  
99 Cas9 cleavage or more easily repaired than others. If this was the case, cells targeted in these regions



**Figure 1. Genome-wide Cas9 killing screen reveals large-scale depletion pattern.** a) A genome-wide library of guide RNA was introduced in *E. coli* strain LC-E19 carrying a *cas9* under the control of a P<sub>tet</sub> promoter. Cells were grown in the presence of 1nM aTc and the guide RNA library sequenced before and after a few hours of induction. b) Scatter plot showing the log<sub>2</sub>FC of guides around the genome. The black line represents the moving average with a window size of 6kb (outer line of circle: log<sub>2</sub>F=2, centre of circle: log<sub>2</sub>F=-6). c) Moving average of guide RNA depletion around the genome after 2H, 4H and 6H of aTc induction. d) Time lapse microscopy after Cas9 induction in the presence of different guide RNAs. e) Fold change of plasmid copy number normalized to a non-targeting control as measured by qPCR. Points show independent biological replicates, the black bar shows the median.

100 might be able to survive longer and possibly keep dividing for a few generations after the induction  
 101 of Cas9. To investigate whether the pattern observed through sequencing was indeed a measure of  
 102 cell survival to Cas9 cleavage, we performed time-lapse microscopy experiments with guide RNA  
 103 targeting different positions either in strongly depleted or weakly depleted regions. In all cases we  
 104 observed that cells started to filament within minutes after Cas9 induction. Cells targeted by weakly  
 105 depleted guides did not show a delayed or weaker response to Cas9 induction.

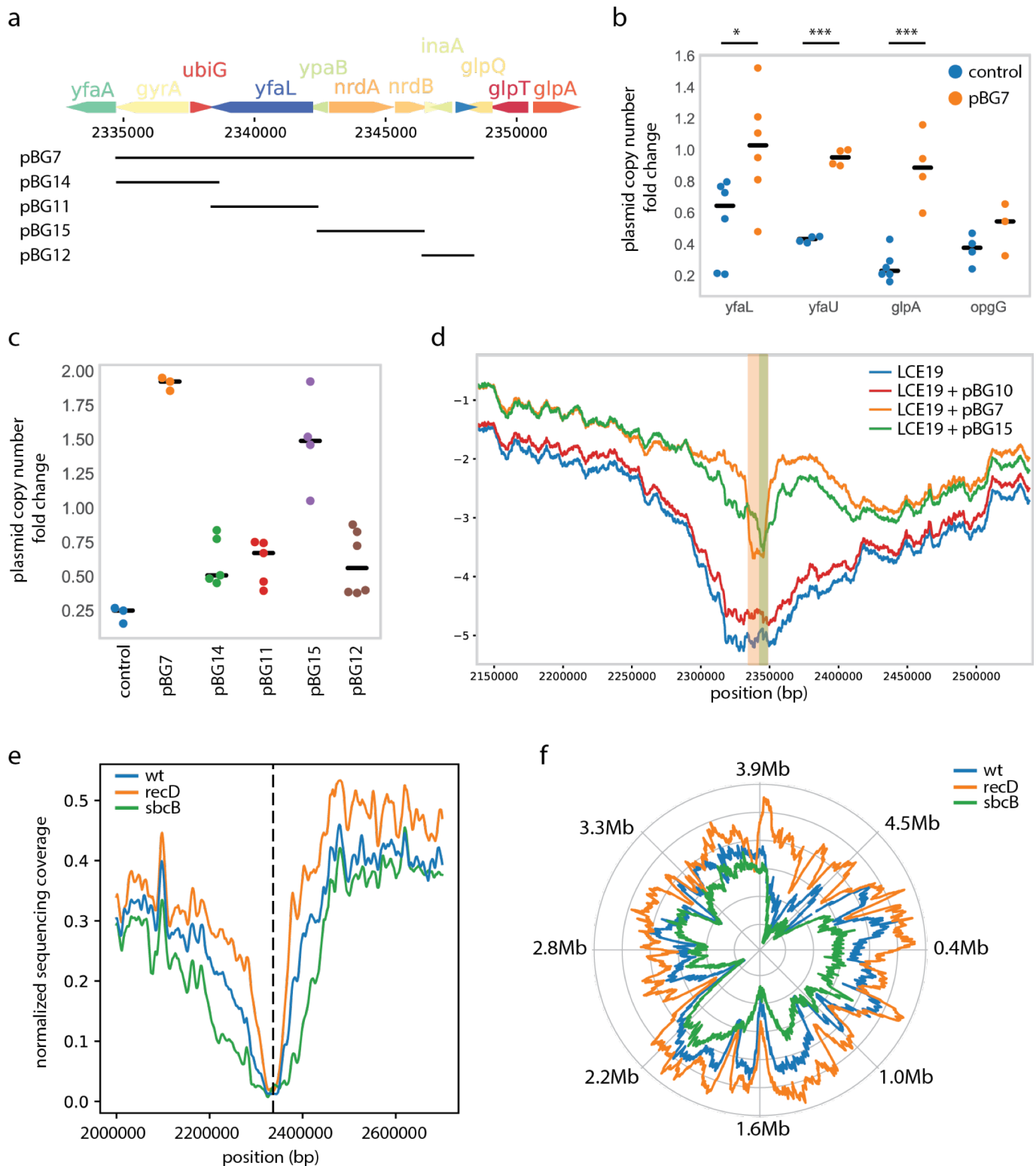
106 The number of reads obtained when sequencing the library can be seen as a measure of the  
 107 abundance of plasmids carrying each guide RNA in the sample, which is typically used as a proxy of  
 108 the number of cells carrying each guide in pooled CRISPR screen assays. Since the large-scale pattern  
 109 observed did not seem to be linked to differences in cell survival, we reasoned that it might rather be  
 110 due to differences in plasmid copy number (PCN). To test our hypothesis, we measured changes in  
 111 PCN by qPCR and normalized to the number of copies of the chromosome as measured with qPCR  
 112 probes distant from the cleavage position. We specifically investigated two regions: one centred on

113 position 2.35 Mb in which guides are strongly depleted and one centred on position 2.95Mb in which  
114 guides are weakly depleted. We measured a decrease of PCN after Cas9 induction when targeting  
115 genes *yfaL* and *glpA* in the first region, and an increase of PCN when targeting genes *fucl* and *tcdA* in  
116 the second region (**Fig. 1e**). These results support the hypothesis that variations in the number of  
117 copies of the plasmid carrying the guide RNAs are responsible for the pattern observed.

### 118 **Degradation of chromosomal genes that control plasmid replication explains the large-scale** 119 **pattern**

120 We then sought to understand why targeting specific genomic regions affects plasmid copy number.  
121 We hypothesized that some regions might contain genes necessary for efficient plasmid replication.  
122 Targeting these regions with Cas9 will lead to the degradation of the target DNA and therefore genes  
123 in the target region will no longer be expressed. If our hypothesis is correct then it should be possible  
124 to restore plasmid replication by cloning genes in the target region on a plasmid that will not be  
125 targeted by Cas9. To test our hypothesis we cloned ~13.3 Kbp located in the central part of the peak  
126 located around 2.34Mbp in plasmid pBG7 (**Fig. 2a**). We then targeted different positions inside or  
127 just outside of the region and measured PCN after Cas9 induction. As expected, the copy number of  
128 the plasmid carrying the guide was higher in the presence of pBG7 than with a control empty vector  
129 (pBG10). When a different region in which guides are also strongly depleted was targeted (see target  
130 in *opgG*), pBG7 did not have an effect on the PCN of psgRNA (**Fig. 2b**). These results support our  
131 hypothesis that genes cloned on pBG7 are important for replication of the plasmid carrying the  
132 library and that the loss of these genes after Cas9 cleavage results in the interruption of plasmid  
133 replication. In order to identify the specific gene responsible for the variations in plasmid copy  
134 number, we independently cloned genes or operons present in the 13.3 Kbp region cloned on pBG7  
135 (**Fig. 2a**). We only observed a substantial increase in PCN when the *nrdA-nrdB-yfaE* operon was  
136 cloned on the plasmid (**Fig. 2c**, see pBG15). Genes *nrdAB* encode a ribonucleoside diphosphate  
137 reductase involved in pyrimidine deoxyribonucleotides synthesis (Kolberg et al. 2004). The loss of  
138 these genes might lead to a decreased availability of deoxypyrimidines and consequently an  
139 inhibition of DNA replication. This hypothesis is also consistent with the fact that the level of *nrdA*  
140 mRNA strongly dropped 4H after Cas9 cleavage, consistently with the emergence of the pattern  
141 (**Supplementary Fig. 1**).

142



**Figure 2. The large-scale pattern is explained by the degradation of chromosomal regions carrying genes necessary for replication.** a) Genes at the centre of a depletion peak were cloned on a set of vectors. b) Fold change in the copy number of the plasmid carrying the guide RNA in the presence of plasmid pBG7 or the control empty vector pBG10, after cleavage with guides targeting genes *yfaL*, *yfaU*, *glpA* in the region, and *opgG* outside of the region. c) Fold change in the copy number of the plasmid carrying the guide RNA in the presence of various complementation plasmids and the control (pBG10), after 4H of induction with a guide targeting *glpA*. d) Moving average of guide RNA depletion in the genome-wide Cas9 killing screen performed in the presence of various complementation plasmids (window size of 6kb). e) Sequence coverage after 30min of Cas9 induction to cut position *yfaL* in strain LC-E19 (wt), a  $\Delta recD$  mutant and a  $\Delta sbcB$  mutant (moving average with a window size of 10kb). f) Moving average of guide RNA depletion in the genome-wide Cas9 killing screen performed in  $\Delta recD$  and  $\Delta sbcB$  mutants (window size of 6kb).



144 Finally, we performed the Cas9 genome-wide screen again in the presence of plasmid pBG7, pBG15  
145 or the control empty vector. We observed a much weaker depletion of guides targeting the region  
146 located around 2.34Mbp in the presence of pBG7 and pBG15 than with the control vector (**Fig. 2d**).  
147 Interestingly guides in the library that targeted the region cloned on the complementation plasmid  
148 showed a stronger depletion than guides targeting just outside of this region (see Supplementary Fig.  
149 2 for a zoom in the region). This is consistent with the fact that these guides will destroy both the  
150 chromosomal and plasmidic *nrdAB* operons, preventing rescue. As an internal control, a mutation  
151 was introduced in the PAM motif of one of the target positions carried by plasmid pBG7. As expected  
152 the guide targeting this position was much less depleted than surrounding guides (**Supplementary**  
153 **Fig. 2**).

#### 154 **The shape of the large-scale pattern is determined by DNA degradation speed**

155 Our results therefore indicate that the *nrdAB* genes might be responsible for the depletion of guides  
156 targeting a region of ~100 Kbp with guides further away from the genes being on average less  
157 depleted than guides closer to the genes. These results suggest that the extent of DNA degradation  
158 determines the shape of the depletion pattern. To investigate this in more details we programmed  
159 plasmid psgRNA to target gene *yfaL* (next to *nrdAB*) and sequenced the DNA extracted from strain  
160 LC-E19 after 30min of induction. Sequence coverage can be used as a measure of the abundance of  
161 DNA around the target region in the sample. DNA was degraded over ~50 Kbp away from the target  
162 in about half of the cells (**Fig. 2e**). The scale of DNA degradation matches that of the large-scale guide  
163 RNA depletion pattern consistently with our hypothesis.

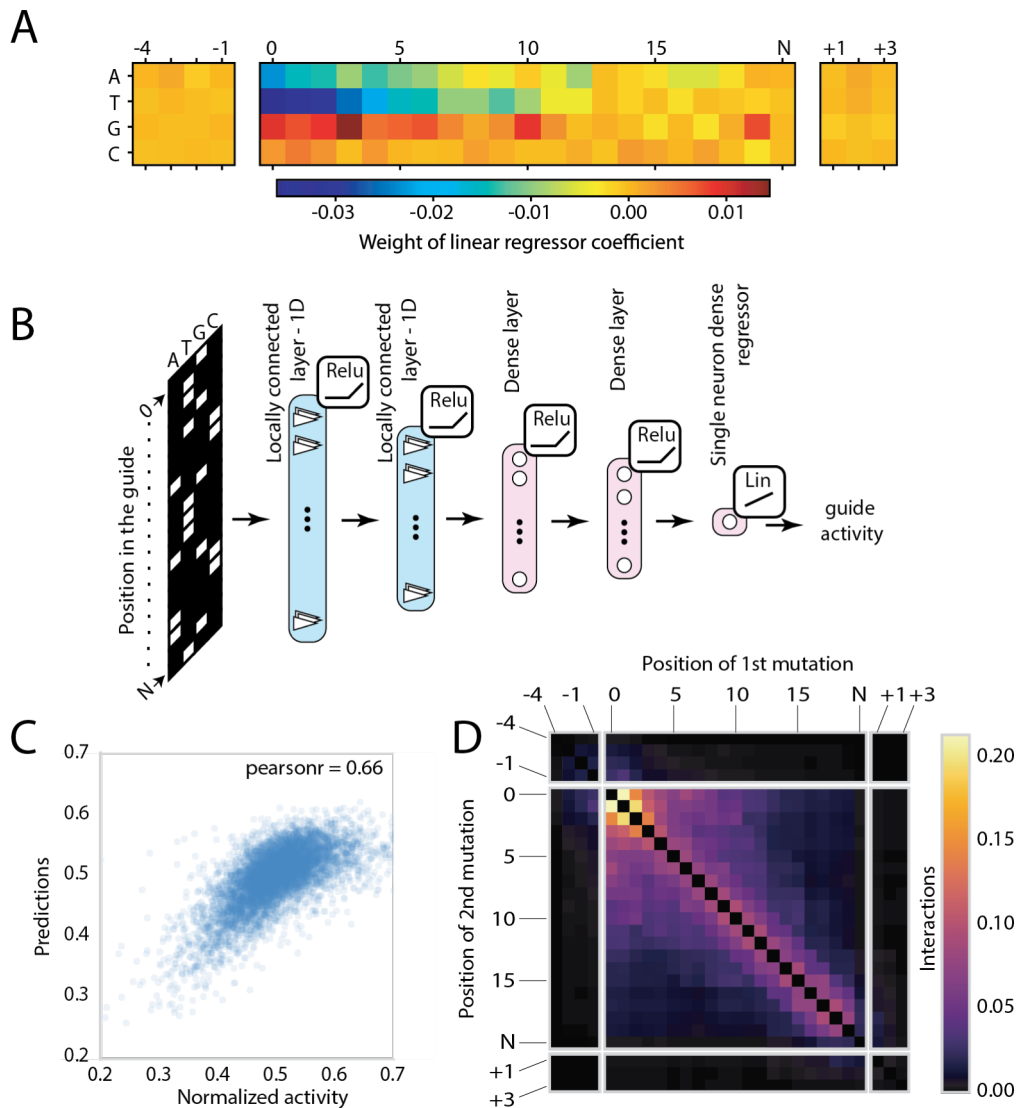
164 If DNA degradation determines the shape of the guide RNA depletion pattern, mutants of genes  
165 involved in DNA degradation might change it. We repeated the whole genome Cas9 cleavage screen  
166 in  $\Delta recD$  and  $\Delta sbcB$  mutants. The pattern was indeed strongly affected in these mutants with  
167 broader peaks and a stronger overall depletion of guides in the  $\Delta sbcB$  mutant and narrower peaks  
168 with a weaker overall depletion of guides in the  $\Delta recD$  mutant (**Fig. 2f**). Consistently with this  
169 observation, DNA degradation was faster in the  $\Delta sbcB$  mutant and slower in the  $\Delta recD$  mutant.  
170 Altogether, our results demonstrate that the large-scale pattern of guide RNA depletion originates  
171 from the effect of degrading some chromosomal regions on plasmid replication and copy number.

#### 172 **A neural network model enables to predict guide RNA efficiency**

173 The large-scale pattern described above accounts for an important part of the variation in the effect  
174 of guides, but we also observed a reproducible variability between guides targeting nearby positions.

175





**Figure 3. Predicting guide RNA efficiency.** (A) A linear model with L1 regularization was trained to predict guide activity using the primary sequence as the only input feature. The heat map shows the coefficients attributed to each base. Position 0 is the first base of the guide, position N refers to the undetermined base of the PAM (NGG), position +1 refers to the first base after the PAM. (B) Architecture of the locally connected neural network trained to predict guide activity from the one-hot-encoded sequence. (C) Predictions of the model plotted as a function of the measured guide activity on a held-out test set. (D) Level of interaction between positions along the target that is seen by the model.

176 In order to eliminate the large-scale effect of plasmid copy number, we computed for each guide in  
 177 the library an activity score as the difference between log<sub>2</sub>FC and the mean log<sub>2</sub>FC of all the guides  
 178 within 6 kbp. The sgRNA activities were extremely consistent across experiments, showing on  
 179 average Pearson correlation between replicates of 0.92 (Supplementary Fig 3). We then sought to  
 180 model guide activity based on the one-hot-encoded target sequence (4 bases upstream, target, PAM  
 181 and 3 bases downstream). A simple linear model with L1 regularization was able to predict guide  
 182 activity in cross-validation with a Pearson correlation of 0.58. The coefficient of the features in the  
 183 regression are plotted as a heat map in Figure 3a. The model uses positions in the guide RNA but not

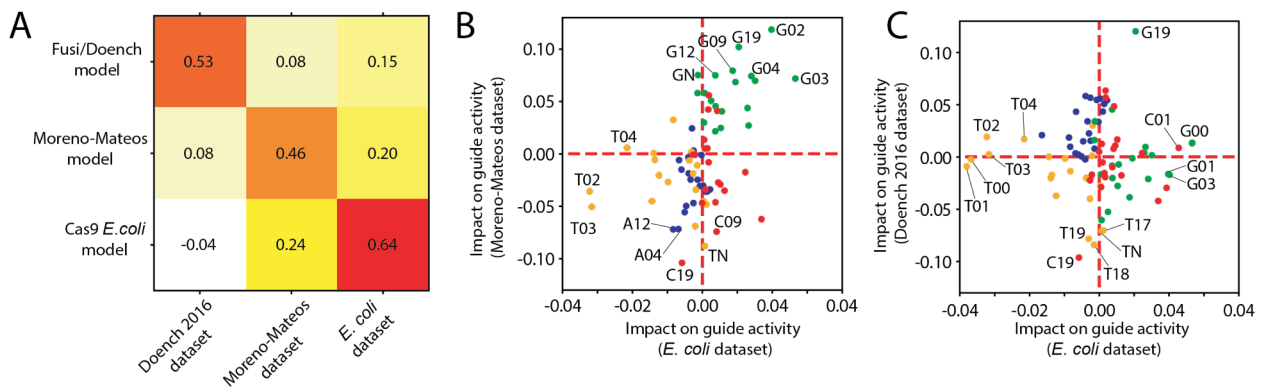
184 the surrounding region in the target. Interestingly bases at the 5' end of the guide RNA impacted the  
185 model predictions the most. The presence of thymidine and to lesser extent adenines seems to  
186 reduce the efficiency of the guides, while guanines and to a lesser extent cytidines increase it. This  
187 observation differs from previous observations in mammalian cells where the PAM-proximal region  
188 was more important than the distal region (H. Xu et al. 2015; Moreno-Mateos et al. 2015; Doench et  
189 al. 2014). Conversely, the presence of a guanine at the last position of the guide has a positive effect  
190 in our model, which was also consistently reported in previous studies.

191 The ability of a simple linear model to obtain good predictions suggests that the determinants of  
192 guide efficiency in *E. coli* are quite simple and can for the most part be inferred from the specific  
193 bases present at each position. This model is however not able to encode more complex interactions  
194 between positions. The large amount of data collected in this experiments enables to train neural  
195 network models which have recently been successfully employed to model sequence data (Kim et al.  
196 2018; Alipanahi et al. 2015). The dataset was split into a training, validation and test set. We first  
197 used a locally connected neural network to predict guide activity using an arbitrary 60nt window  
198 around the target (Figure 3b). The resulting model achieved a Pearson correlation coefficient of 0.66  
199 on the held-out test set (Spearman-r=0.64). In order to investigate the sequence features used by the  
200 model to make its predictions we generated a set of 1000 random sequences and measured the  
201 impact on the model predictions of mutating each position to all possible bases. The results are  
202 consistent with the linear regression model, showing the importance of the 5' end of the guide, while  
203 giving little to no importance to positions outside of the 20nt target (Supplementary Figure 4). We  
204 thus restricted the inputs of the model to target sequence, 4 bases upstream and 3 bases  
205 downstream of the PAM. This simpler model still achieved a Pearson correlation of 0.66 on the test  
206 set while reducing overfitting on the training set (Figure 3c).

207 In order to investigate the interactions identified by the model we further mutated each pair of  
208 position to all possible bases for a set of 200 random sequences. We then compared the effect of a  
209 double mutation and the sum of the effect the two single mutations. It appears clearly that bases in  
210 the guide interact with their direct neighbours, and to a lesser extent with their second and even  
211 third neighbours (Figure 3d). Interestingly positions in the first half of the guide seem to interact with  
212 each other on a long range, an effect not observed in the PAM-proximal region.

### 213 **Comparison to other models**

214 We then compared our model to two of the most widely used models for on-target activity  
215 prediction. Doench and colleagues modeled the activity of guide RNAs expressed from a U6  
216 promoter on a DNA vector in human cells (Doench et al. 2016), while Moreno-Mateos and colleagues



**Figure 4. Comparison between models.** (A) Heat map of Spearman correlation coefficient between activity scores predicted by models and datasets. The score given for the performance of models on their own training set corresponds to the performances on a test set in cross-validation. The other correlations are computed on the whole dataset. (B, C) Comparison of the impact of specific bases along the guide sequence between the different datasets, i.e. the difference between the average activity of guides that have this feature and the average activity of guides that do not have it. The following color code is used for bases (Adenine: blue, Thymine: yellow, Guanine: green, Cytosine: red).

217 modeled the activity of guide RNAs expressed *in vitro* from a T7 promoter and injected into zebrafish  
 218 embryo (Moreno-Mateos et al. 2015).

219 A previous report showed that these two models make very different predictions suggesting that the  
 220 determinants of guide efficiency strongly depend on the exact experimental setup (Haeussler et al.  
 221 2016). When applied to our data, the model of Moreno-Mateos provides some predictive power, and  
 222 conversely our model shows some predictive power on the data from Moreno-Mateos. The model of  
 223 Doench performs weak predictions on the *E. coli* and zebrafish data. In order to shed light on the  
 224 specific sequence features that matter in these different datasets, we plotted the impact of having  
 225 specific bases at each position along the guide RNA, i.e. the difference between the average activity  
 226 of guides that have this feature and the average activity of guides that do not have it. This analysis  
 227 reveals that many of the features that impact the activity of guides in *E. coli* also impact the activity  
 228 of guides in the zebrafish dataset. Note that because of experimental constraints, the dataset of  
 229 Moreno-Mateos only includes guides whose first two bases are guanines. In our dataset these two  
 230 bases are the most important features determining guide activity, but their effect is unknown in the  
 231 experimental setup of Moreno-Mateos.

232

## 233 Discussion

234 Over the last few years numerous studies have led to a detailed understanding of Cas9 target search  
 235 mechanism, structure, and biochemical activity (Gasiunas et al. 2012; Jinek et al. 2012; S. H.

236 Sternberg et al. 2014; Jinek et al. 2014; Szczelkun et al. 2014; Anders et al. 2014; Samuel H. Sternberg  
237 et al. 2015). Despite the impressive amount of knowledge gathered, little is known about how the  
238 guide RNA sequence affects the efficiency of Cas9 targeting. Several groups have performed high-  
239 throughput screens in eukaryotic cells to measure the activity of thousands of guide RNAs, enabling  
240 the development of algorithms that can predict the efficiency with which a guide RNA can knockout a  
241 gene (Doench et al. 2014; Moreno-Mateos et al. 2015; H. Xu et al. 2015; Doench et al. 2016).  
242 Strikingly, these algorithms tend to perform well on some datasets and very poorly on others  
243 (Haeussler et al. 2016; Labuhn et al. 2018). Some of these differences have been attributed to the  
244 specificities of the experimental system and whether the guide RNA is transcribed *in vitro* from a T7  
245 promoter followed by injection in cells, or *in vivo* from a U6 promoter on a DNA vector (Haeussler et  
246 al. 2016). For instance, assays based on DNA vectors consistently observed a bias against uracil at the  
247 3' end of the guide RNA sequence, which is due to the propensity of the RNA polymerase III to  
248 terminate transcription at uridine-rich sequences (Doench et al. 2014; Wang et al. 2014; Wu et al.  
249 2014; Moreno-Mateos et al. 2015). These results argue for the necessity to develop guide RNA design  
250 tools that are specific to each experimental setup.

251 Here we measured guide RNA depletion after cleavage in the chromosome of *E. coli*, enabling the  
252 investigation of features that affect Cas9 targeting in a completely different experimental setup than  
253 previous reports, and with a lot more data points. In this assay efficient guides kill the bacteria and  
254 are rapidly depleted. The most striking feature that explained differences between guide RNAs was  
255 their position along the chromosome. Studies in Eukaryotic systems have shown the importance of  
256 chromatin structure and accessibility on Cas9 targeting efficiency (Kuscu et al. 2014; Wu et al. 2014;  
257 Horlbeck et al. 2016; Chen et al. 2016). However, the pattern of guide RNA depletion that we  
258 observed here did not match the known structure of the *E. coli* chromosome (Liyo et al. 2018).  
259 Rather, we were able to demonstrate that this large-scale pattern emerges from changes in plasmid  
260 copy number that occur after the degradation of specific chromosome regions after Cas9 cleavage.  
261 The description of this artefact should serve as a cautionary tale for any pooled genetic screen in  
262 which the readout is made by sequencing a guide RNA or any barcode on a plasmid. Note that Cas9  
263 might still be able to kill *E. coli* more or less efficiently depending on which region of the  
264 chromosome is targeted, but a different experimental design will have to be used to investigate this  
265 question.

266 After taking the large-scale depletion pattern into account we could investigate the effect of the  
267 guide RNA sequence on the ability of Cas9 to kill *E. coli*. Our results highlighted the positive effect of  
268 a high GC content at the 5' end of the guide RNA, and in particular the negative effect of thymidines  
269 in this region. Binding of the 5' end to the target sequence is required for a conformational shift to

270 occur in Cas9 which brings the RuvC and HNH catalytic domains in contact with the target DNA  
271 (Samuel H. Sternberg et al. 2015; Jinek et al. 2014). This binding might be disfavored by a high  
272 thymidine content, making cleavage less efficient. Another possible explanation is that guide RNA  
273 transcription might be imperfect for some sequences. It was previously reported that the sequence  
274 around the transcriptional start site (TSS) can impact the frequency at which the polymerase will  
275 initiate transcription at different positions (Vvedenskaya et al. 2015). In particular, thymidine-rich  
276 sequences were shown to favor initiation further away from the -10 element, which in our case  
277 would lead to the formation of truncated guide RNAs. The same study also highlighted the fact that  
278 poly-T and poly-A stretches after the TSS promote slippage synthesis, which could conversely lead to  
279 the formation of longer guide RNAs. Finally, the sequence at the 5' end of the guide might impact its  
280 stability *in vivo*. Moreno-Mateos measured an increased stability and activity of guides with a higher  
281 guanine content, and provided evidence that these guides might form G-quadruplexes protecting  
282 them against 5'-directed exonucleases. This mechanism could also be at play in our experiments.

283 Altogether, our results provide insight into the features that determine the ability of Cas9 to  
284 efficiently kill *E. coli*. We provide a model able to predict the most efficient guides with a high  
285 accuracy that should be useful for the design of effective guides in genome editing experiments as  
286 well as for applications of CRISPR as antimicrobials (Jiang et al. 2013; Cui and Bikard 2016; Bikard et  
287 al. 2014; Citorik, Mimee, and Lu 2014). This model is made available online:

288 <http://hub13.hosting.pasteur.fr:8080/CRISPRBact/>

289

## 290 **Methods**

### 291 **Bacterial strains and media**

292 *E. coli* strains were grown in Luria-Bertani (LB) broth or LB Agar 1.5% as solid medium. Whenever  
293 applicable, media was supplemented with chloramphenicol ( $20 \mu\text{g ml}^{-1}$ ), carbenicillin ( $100 \mu\text{g ml}^{-1}$ ) or  
294 kanamycin ( $50 \mu\text{g ml}^{-1}$ ) to select or ensure the maintenance of the plasmids. Lower concentration of  
295 kanamycin ( $20 \mu\text{g ml}^{-1}$ ) was used to select for the integration of vectors in the chromosome. All the  
296 strains modifications derived from *E. coli* MG1655. *E. coli* strain DH5 $\alpha$  or MG1655 were used as  
297 transformation recipients.

### 298 **Plasmids and *E. coli* strains construction**

299 Strain LC-E19 was constructed using the pOSIP system (St-Pierre et al. 2013) and the backbones were  
300 removed using the pE-FLP plasmid (St-Pierre et al. 2013). The Ptet-Cas9 expression cassette was

301 integrated at the HK022 *attB* and a *gfp* (Green Fluorescent Protein) reporter gene under the control  
302 the *sulA* promoter was integrated at  $\lambda$  *attB* to monitor SOS response. Genes *recD* and *sbcB* were  
303 deleted from strain LC-E19 using the lambda red recombineering strategy (Sharan et al. 2009).  
304 Plasmid pKD4 was used as a template to generate linear DNA fragments via polymerase chain  
305 reaction (PCR) followed by electroporation into strain LC-E19 carrying plasmid pKOBEG-A  
306 (Chaverroche, Ghigo, and d' Enfert 2000). Colonies resistant to kanamycin were selected and the  
307 resistance gene was then removed using plasmid pE-FLP. The constructions were verified by PCR and  
308 sequencing. All the bacterial strains used in this study are listed in Supplementary Table 1 and the  
309 primers used for strain construction in Supplementary Table 2.

310 Fragments for plasmid constructions were generated by PCR or restriction digestion and assembled  
311 through Gibson assembly (Gibson et al. 2009). Novel guide RNAs were cloned into plasmid psgRNA by  
312 golden gate assembly (Engler, Kandzia, and Marillonnet 2008). A list of plasmids is provided as  
313 Supplementary Table 3, and primers used in plasmid construction as Supplementary Table 4. A list of  
314 guide RNA used in the study is provided as Supplementary Table 5.

### 315 **Genome-wide Cas9 screen**

316 The data shown in figure 1 was generated as follow. The guide RNA library carried on plasmid  
317 psgRNA was electroporated into LC-E19, plated in LB with kanamycin 50  $\mu$ g/ml and incubated at 37°C  
318 for 4h. An estimated number of  $10^7$  clones were recovered, pooled in 10ml LB and stored as 1ml  
319 aliquots in DMSO 10% at -80°C. To perform the assay, 1ml of frozen cells were thawed in 400ml of LB  
320 with kanamycin 50  $\mu$ g/ml and cultivated at 37°C, 190 RMP until they reached early-exponential phase  
321 ( $OD_{600} \approx 0.25$ ). Then, aTc 1nM was added and cells were recovered at different time points (0h, 2h,  
322 4h and 6h). Plasmids were extracted from 50 ml of culture using the NucleoSpin Plasmid kit  
323 (Macherey-Nagel, Duren, Germany). The whole assay was performed in triplicate.

324 During the course of the study we realized that leaky expression of Cas9 in strain LCE-19 might lead  
325 to the introduction of biases in the library as clones that mutate the CRISPR-Cas9 system or the  
326 target might be selected before induction with aTc. To avoid this problem, we used modified  
327 experimental design for the data shown in figure 2 and figure 3. Plasmid psgRNA carries a *cos* site  
328 enabling its packaging in phage lambda capsids. The library was transformed in strain CY2120 which  
329 carries a temperature sensitive lysogenic lambda prophage with its *cos* site deleted (Cronan 2013).  
330 Upon induction at 42°C, lambda capsids are produced and the psgRNA packaged. Cosmid particles  
331 can then be purified as described previously (Cronan 2013). Briefly, strain CY2120 carrying psgRNA  
332 was diluted 100-fold from an overnight culture in LB supplemented with 50 mM Tris-HCl buffer (pH  
333 7.5) and grown at 30°C until  $OD_{600} \approx 0.5$ . The culture was then incubated at 42°C for 20 minutes

334 and then at 37°C for 4 hours. Cells were harvested by centrifugation at 4000g for 5 minutes, washed  
335 in lambda dilution buffer (Tris-HCL pH 7.5 20mM, NaCl 0.1M, MgSO4 10mM) with 1/20 of the culture  
336 volume, centrifuged and re-suspended again in the same volume of lambda dilution buffer. To purify  
337 the cosmid particles chloroform was added (20 vol%), samples vortexed for 15 seconds and  
338 incubated at 37°C for 15 minutes with shaking. Cosmids were harvested after centrifugation at  
339 12,000g for 1.5 minutes.

340 This cosmid library was then used to transduce strain LC-E19 induced to produce Cas9 as follow. An  
341 overnight culture of strain LC-E19 was diluted 100-fold in fresh LB, 0,2% arabinose and aTc 1nM and  
342 grown until  $OD_{600} \approx 0.6$ . The psgRNA library was transduced at a MOI (Multiplicity of Infection) of 0.2  
343 and incubated at room temperature for 30 minutes. Cells were then grown 4h at 37°C, 190 RPM with  
344 kanamycin 50 µg/ml to inhibit the growth of cells that were not transduced. Cells were recovered  
345 and plasmids extracted from 15 ml of culture using NucleoSpin Plasmid kit (Macherey-Nagel, Duren,  
346 Germany). Transduction in strain MG1655, which does not carry Cas9, was used as a control.

347 To measure the relative abundance of guide RNAs in each sample, the library of guide RNA was  
348 sequenced following the method described by Lun Cui *et al.* (Cui *et al.* 2018). Two nested PCR  
349 reactions were used to generate the sequencing library with primers described in Supplementary  
350 Table 6. The 1<sup>st</sup> PCR adds the 1<sup>st</sup> index. The 2<sup>nd</sup> PCR adds the 2<sup>nd</sup> index and flow cells attachment  
351 sequences. Sequencing is then performed using primer LC609 as a custom read 1 primer. Custom  
352 index primers were also used: LC499 reads index 1 and LC610 reads index 2. Sequencing was  
353 performed on a NextSeq 500 benchtop sequencer. The first 2 sequencing cycles read bases common  
354 to all clusters and were set as dark cycles, followed by 20 cycles corresponding to the guide RNA.

355 We report here the log2 transformed fold change in the number of reads obtained for each guide  
356 RNA, normalized by the total number of reads obtained for the sample and the number of reads  
357 obtained for a control guide RNA which does not have a target position in the genome of *E. coli*  
358 MG1655. Reads were counted only if they showed a perfect match to the sequence of a guide in the  
359 library.

#### 360 **Determination of plasmid copy numbers**

361 Pre-cultures of LCE-19 carrying a plasmid with a single guide RNA were incubated at 37°C in triplicate  
362 at 190 RPM shaking overnight. The appropriate antibiotics (kanamycin 50 µg/ml or chloramphenicol  
363 20 µg/ml) were used to maintain the plasmids. Pre-cultures were diluted 100-fold in fresh LB and  
364 grown in the same conditions until they reached early-exponential phase ( $OD_{600} \approx 0.25$ ). Cas9  
365 expression was then induced with aTc 1nM. Cells were recovered before and after 4 hours of  
366 induction. Total DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega,



367 Madison, USA). Quantitative Polymerase Chain Reaction (qPCR) was carried out in triplicate for each  
368 extraction using the FastStart Essential DNA Green Master (Roche Diagnostics, Mannheim, Germany)  
369 in accordance with the manufacturer's specifications in a LightCycle 96 system (Roche Diagnostics).  
370 The plasmid copy number was determined as described by San Millan *et al.* (San Millan, Heilbron,  
371 and Maclean 2013). The efficiencies of the reactions were calculated from a standard curve  
372 generated by performing qPCR with five 10-fold dilutions of template DNA in triplicate (10 ng/ $\mu$ l-  
373 1pg/ $\mu$ l). Primers BG114 and BG115 were used to perform the qPCR with a concentration of template  
374 DNA of 0.1 ng/ $\mu$ l. To determine the average PCN per chromosome, we used primers BG116 and  
375 BG117 to amplified gene *rpoB*. Primers are listed in Supplementary table 7. The amplification  
376 conditions were: initial denaturation for 10 min at 95°C, followed by 45 cycles of 10 s at 95°C, 20 s at  
377 55°C and 10 s at 72°C. The following formula described by San Millan *et al.* (San Millan, Heilbron, and  
378 Maclean 2013) was used:

379

$$cn = \frac{(1 + E_c)^{C_{tc}}}{(1 + E_p)^{C_{tp}}} * \frac{S_c}{S_p}$$

380 where *cn* is the plasmid copy number per chromosome,  $S_c$  and  $S_p$  are the sizes of the chromosomal  
381 and plasmid amplicons (in bp),  $E_c$  and  $E_p$  are the efficiencies of the chromosomal and plasmid qPCRs  
382 (relative to 1), and  $C_{tc}$  and  $C_{tp}$  are the threshold cycles of the chromosomal and plasmid reactions,  
383 respectively. Fold change values in plasmid copy number after 4H of Cas9 induction and normalized  
384 to a non-targeting control are reported.

### 385 **Quantification of gene expression**

386 Overnight cultures were diluted 100-fold and grown until they reached early-exponential phase  
387 ( $OD_{600} \approx 0.25$ ). Then aTc (1nM) was added and cells incubated for 4h. RNA was extracted from 10 ml  
388 of culture at  $OD_{600} \approx 0.25$ , and from 2ml after 4h of induction using Direct-zol™ RNA Miniprep kit  
389 (Zymo Research). All the RNA samples were first treated with DNase (Turbo DNase free kit, Ambion)  
390 and then 1  $\mu$ g of RNA for each sample was reverse transcribed into cDNA using the Transcriptor First  
391 strand cDNA synthesis Kit (Roche). qPCR was performed using 1  $\mu$ L of the reverse transcription  
392 reaction with the Faststart essential DNA green master mix (Roche) in a LightCycle 96 (Roche). Probes  
393 and PCR primers are listed in Supplementary Table 7. Relative gene expression was computed using  
394 the  $\Delta\Delta Cq$  method (Schmittgen and Livak 2008).

### 395 **Quantification of DNA degradation after Cas9 cleavage**

396 Strain LC-E19 carrying plasmid psgRNA programmed to target gene *yfaL* was diluted 100x from on  
397 overnight culture and grown until  $OD_{600} \approx 0.25$ . Cas9 expression was induced by addition of aTc 1nM  
398 followed by 30min of incubation at 37°C with shaking. Total DNA was extracted using the Wizard  
399 Genomic DNA Purification Kit (Promega, Madison, USA) and fragmented with a Covaris E220  
400 ultrasonicator. Sequencing libraries were prepared using the NEXTflex PCR-free DNA-Seq kit (Bioo  
401 Scientific Corporation), and sequenced on a MiSeq v3 PE300 flowcell. Sequencing reads were  
402 mapped to the genome of *E. coli* LC-E19 using bowtie2 (Langmead and Salzberg 2012) and sequence  
403 coverage computed with samtools (Li et al. 2009). The plot of figure 2e shows a moving average of  
404 the coverage normalized between 0 and 1, with a window of 10kb.

#### 405 **Microscopy**

406 LC-E19 cells were transformed with plasmid psgRNA carrying different guides. Overnight cultures  
407 were diluted 100-fold in fresh LB and incubated at 37°C for 1h. Cas9 expression was induced with aTc  
408 1nM and cells transferred to LB pads with 1% UltraPure™ agarose (Invitrogen), kanamycin 50 µg/ml  
409 and aTc 1nM. Cells were imaged using an inverted microscope (TI-E, Nikon Inc.) equipped with a 100x  
410 phase contrast objective (CFI PlanApo Lambda DM100x 1.4NA, Nikon Inc.). Images were taken every  
411 5 min during 8 h with an exposure of 100 ms using a sCMOS camera (Orca Flash 4.0, Hamamatsu)  
412 with an effective pixel size of 65 nm.

#### 413 **Machine learning**

414 For each guide of the library, we computed a sgRNA activity score as the difference between its  
415 log2FC and the mean log2FC of all the guides within 6 Kbp, normalized between 0 and 1. The activity  
416 score of guides was computed as the mean activity from 4 independent assays measured in strain LC-  
417 E19 using the cosmid transduction assay.

418 After splitting the dataset into a training, validation and test sets, we implemented a neural network  
419 (Fig 3b) using Keras and TensorFlow. We used a model architecture inspired from our previous work  
420 (Cui et al. 2018) and consisting of 2 locally-connected layers of size 20 and 8, with kernel sizes of 4  
421 and 7, followed by 2 dense layers of size 14 and 8. The layers sizes and kernel sizes were optimized to  
422 minimize the loss on the validation set using a grid search approach. We used the rectified linear unit  
423 (ReLU) activation function for all the neurons of these layers. Finally a densely connected single  
424 neuron predicted the sgRNA efficacy using a linear combination of the last dense layer. The network  
425 was trained to minimize the mean square error of the log2FC prediction with L2 regularization using  
426 the Adam optimizer (Kingma and Ba 2014). Training was interrupted when loss on the validation set  
427 ceased to decrease for more than two epochs.

428 To identify the positions used by the model to make its predictions we generated a set of 1000  
429 random sequences, mutated each position *in silico*, and computed the effect of each mutation on the  
430 model prediction (Supplementary Figure 4). To measure the level of interaction between positions  
431 we generated all possible pairs of mutations for each sequence in a set of 200 random sequences,  
432 and compared the effect of individual mutations to that of pairs of mutations. Positions are  
433 interacting if the effect of a double mutation ( $E_{ij}$ ) is different from the sum of the effect of the single  
434 mutations ( $E_i+E_j$ ). The heat map shows the average Euclidean distance between  $E_{ij}$  and  $E_i+E_j$  for all  
435 pairs of positions (Figure 3).

### 436 **Model comparisons**

437 The Fusi/Doench algorithm (from Doench 2016) was recoded from explanations provided in the  
438 article. In order to have a fair comparison with our model, all the features detailed in Doench 2016  
439 were used except the cutting position and the amino acid percentage position. We then performed a  
440 5-fold cross validation using a boosted regression tree on the FC+RES dataset from the same paper  
441 and obtained a Spearman coefficient similar to what was reported. In order to test the prediction  
442 capabilities of the Fusi/Doench algorithm on other datasets, we re-trained the algorithm on the  
443 whole FC+RES dataset and computed the correlation between the model predictions and the actual  
444 sgRNA activity.

445 Similarly, we recoded the Moreno-Mateos algorithm. Features far from the guide sequence that  
446 were used in the Moreno-Mateos were not kept as their value was not easily accessible for the  
447 FC+RES data. Removing them did not hurt the predictions. We then used a lassoCV regression and  
448 performed a 5-fold cross validation on the Moreno-Mateos dataset (obtained from Haeussler and  
449 colleagues (Haeussler et al. 2016)). We obtained similar prediction performance than described in  
450 their original paper (Spearman coefficient = 0.46). In order to test the prediction capabilities of the  
451 Moreno-Mateos algorithm on other datasets, we retrained the model on the whole Moreno-Mateos  
452 dataset and computed the correlation between the model predictions and the actual sgRNA activity.

### 453 **Data Access**

454 The screen results are provided as Supplementary Dataset 1 and Supplementary Dataset 2. Other  
455 relevant data supporting the findings of the study are available in this article and its Supplementary  
456 Information files, or from the corresponding author upon request. An online version of the model is  
457 provided here: <http://hub13.hosting.pasteur.fr:8080/CRISPRBact/>

### 458 **Acknowledgements**

459 We thank Gizem Ozbaykal for the help with the fluorescence and phase-contrast microscopy. DNA  
460 sequencing after Cas9 cleavage was performed by Cédric Fund from the Institut Pasteur Genomics  
461 Platform, a member of “France Génomique” consortium (ANR10-INBS-09-08). This work was  
462 supported by the European Research Council (ERC) under the Europe Union’s Horizon 2020 research  
463 and innovation program (grant agreement No [677823]); the French Government's Investissement  
464 d'Avenir program; Laboratoire d'Excellence ‘Integrative Biology of Emerging Infectious Diseases’  
465 [ANR-10-LABX-62-IBEID]; the Pasteur-Weizmann consortium, and a Ramon Areces Fellowship to B.G.

## 466 **Author contributions**

467 B.G, L.C, J.W.N and D.B. designed the study and wrote the manuscript. B.G. and L.C performed the  
468 experiments. C.B. developed CRISPRbact. D.B. and J.W.N analysed the data. B.G, D.B. and J.W.N.  
469 wrote the manuscript.

470

## 471 References

- 472 Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the  
473 Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." *Nature*  
474 *Biotechnology* 33 (8): 831–38. <https://doi.org/10.1038/nbt.3300>.
- 475 Anders, C., O. Niewoehner, A. Duerst, and M. Jinek. 2014. "Structural Basis of PAM-Dependent Target  
476 DNA Recognition by the Cas9 Endonuclease." *Nature*, July.  
477 <https://doi.org/10.1038/nature13579>.
- 478 Ayora, Silvia, Begoña Carrasco, Paula P. Cárdenas, Carolina E. César, Cristina Cañas, Tribhuwan Yadav,  
479 Chiara Marchisone, and Juan C. Alonso. 2011. "Double-Strand Break Repair in Bacteria: A  
480 View from *Bacillus Subtilis*." *FEMS Microbiology Reviews* 35 (6): 1055–81.  
481 <https://doi.org/10.1111/j.1574-6976.2011.00272.x>.
- 482 Bernheim, Aude, Alicia Calvo-Villamañán, Clovis Basier, Lun Cui, Eduardo P. C. Rocha, Marie Touchon,  
483 and David Bikard. 2017. "Inhibition of NHEJ Repair by Type II-A CRISPR-Cas Systems in  
484 Bacteria." *Nature Communications* 8 (1): 2094. <https://doi.org/10.1038/s41467-017-02350-1>.
- 485 Bikard, David, Chad W Euler, Wenyan Jiang, Philip M Nussenzweig, Gregory W Goldberg, Xavier  
486 Duportet, Vincent A Fischetti, and Luciano A Marraffini. 2014. "Exploiting CRISPR-Cas  
487 Nucleases to Produce Sequence-Specific Antimicrobials." *Nat Biotech* 32 (11). Nature  
488 Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 1146–50.
- 489 Bikard, David, Asma Hatoum-Aslan, Daniel Mucida, and Luciano A Marraffini. 2012. "CRISPR  
490 Interference Can Prevent Natural Transformation and Virulence Acquisition during in Vivo  
491 Bacterial Infection." *Cell Host & Microbe* 12 (2). Elsevier: 177–86.  
492 <https://doi.org/10.1016/j.chom.2012.06.003>.
- 493 Chaverroche, Marie-Kim, Jean-Marc Ghigo, and Christophe d' Enfert. 2000. "A Rapid Method for  
494 Efficient Gene Replacement in the Filamentous Fungus *Aspergillus Nidulans*." *Nucleic Acids*  
495 *Research* 28 (22): e97.
- 496 Chen, Xiaoyu, Marrit Rinsma, Josephine M. Janssen, Jin Liu, Ignazio Maggio, and Manuel A.F.V.  
497 Gonçalves. 2016. "Probing the Impact of Chromatin Conformation on Genome Editing Tools."  
498 *Nucleic Acids Research* 44 (13): 6482–92. <https://doi.org/10.1093/nar/gkw524>.
- 499 Citorik, R. J., M. Mimee, and T. K. Lu. 2014. "Sequence-Specific Antimicrobials Using Efficiently  
500 Delivered RNA-Guided Nucleases." *Nat Biotechnol* 32 (11): 1141–45.  
501 <https://doi.org/10.1038/nbt.3011>.
- 502 Cronan, John E. 2013. "Improved Plasmid-Based System for Fully Regulated off-to-on Gene  
503 Expression in *Escherichia Coli*: Application to Production of Toxic Proteins." *Plasmid* 69 (1):  
504 81–89. <https://doi.org/10.1016/j.plasmid.2012.09.003>.
- 505 Cui, Lun, and David Bikard. 2016. "Consequences of Cas9 Cleavage in the Chromosome of *Escherichia*  
506 *Coli*." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkw223>.
- 507 Cui, Lun, Vigouroux, Antoine, Francois Rousset, Hugo Varet, Varun Khanna, and David Bikard. 2018.  
508 "A CRISPRi Screen in *E. Coli* Reveals a Sequence-Specific Toxicity of DCas9." *Nature*  
509 *Communications* in press.
- 510 Dillingham, M. S., and S. C. Kowalczykowski. 2008. "RecBCD Enzyme and the Repair of Double-  
511 Stranded DNA Breaks." *Microbiology and Molecular Biology Reviews*.  
512 <https://doi.org/10.1128/MMBR.00020-08>.
- 513 Doench, J. G., Nicolo Fusì, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F.  
514 Donovan, Ian Smith, et al. 2016. "Optimized SgRNA Design to Maximize Activity and Minimize  
515 Off-Target Effects of CRISPR-Cas9." *Nature Biotechnology* 34 (2): 184–91.  
516 <https://doi.org/10.1038/nbt.3437>.
- 517 Doench, J. G., E. Hartenian, D. B. Graham, Z. Tothova, M. Hegde, I. Smith, M. Sullender, B. L. Ebert, R.  
518 J. Xavier, and D. E. Root. 2014. "Rational Design of Highly Active SgRNAs for CRISPR-Cas9-  
519 Mediated Gene Inactivation." *Nat Biotechnol* 32 (12): 1262–67.  
520 <https://doi.org/10.1038/nbt.3026>.

- 521 Engler, Carola, Romy Kandzia, and Sylvestre Marillonnet. 2008. "A One Pot, One Step, Precision  
522 Cloning Method with High Throughput Capability." *PLoS ONE* 3 (11).  
523 <https://doi.org/10.1371/journal.pone.0003647>.
- 524 Gasiunas, Giedrius, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. 2012. "Cas9–crRNA  
525 Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in  
526 Bacteria." *Proceedings of the National Academy of Sciences* 109 (39): 15539–40.  
527 <https://doi.org/10.1073/pnas.1208507109>.
- 528 Gibson, Daniel G, Lei Young, Ray-Yuan Chuang, J Craig Venter, Clyde A Hutchison III, and Hamilton O  
529 Smith. 2009. "Enzymatic Assembly of DNA Molecules up to Several Hundred Kilobases."  
530 *Nature Methods* 6 (April). Nature Publishing Group: 343.
- 531 Gomaa, A. A., H. E. Klumpe, M. L. Luo, K. Selle, R. Barrangou, and C. L. Beisel. 2013. "Programmable  
532 Removal of Bacterial Strains by Use of Genome-Targeting CRISPR-Cas Systems." *MBio* 5 (1):  
533 e00928-13. <https://doi.org/10.1128/mBio.00928-13>.
- 534 Haeussler, Maximilian, Kai Schönig, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste  
535 Renaud, Sylvie Schneider-Maunoury, et al. 2016. "Evaluation of Off-Target and on-Target  
536 Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR." *Genome*  
537 *Biology* 17 (July): 148. <https://doi.org/10.1186/s13059-016-1012-2>.
- 538 Horlbeck, Max A., Lea B. Witkowsky, Benjamin Guglielmi, Joseph M. Replogle, Luke A. Gilbert,  
539 Jacqueline E. Villalta, Sharon E. Torigoe, Robert Tjian, and Jonathan S. Weissman. 2016.  
540 "Nucleosomes Impede Cas9 Access to DNA in Vivo and in Vitro." *ELife* 5 (March): e12677.  
541 <https://doi.org/10.7554/eLife.12677>.
- 542 Hsu, P. D., E. S. Lander, and F. Zhang. 2014. "Development and Applications of CRISPR-Cas9 for  
543 Genome Engineering." *Cell* 157 (6): 1262–78. <https://doi.org/10.1016/j.cell.2014.05.010>.
- 544 Jiang, W., D. Bikard, D. Cox, F. Zhang, and L. A. Marraffini. 2013. "RNA-Guided Editing of Bacterial  
545 Genomes Using CRISPR-Cas Systems." *Nat Biotechnol* 31 (3): 233–39.  
546 <https://doi.org/10.1038/nbt.2508>.
- 547 Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A Programmable  
548 Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* 337 (6096):  
549 816–21. <https://doi.org/10.1126/science.1225829>.
- 550 Jinek, M., F. Jiang, D. W. Taylor, S. H. Sternberg, E. Kaya, E. Ma, C. Anders, et al. 2014. "Structures of  
551 Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation." *Science* 343 (6176):  
552 1247997. <https://doi.org/10.1126/science.1247997>.
- 553 Kim, Hui Kwon, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim,  
554 Sangeun Lee, Sungroh Yoon, and Hyongbum (Henry) Kim. 2018. "Deep Learning Improves  
555 Prediction of CRISPR–Cpf1 Guide RNA Activity." *Nature Biotechnology* 36 (3): 239–41.  
556 <https://doi.org/10.1038/nbt.4061>.
- 557 Kingma, Diederik P., and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization."  
558 *ArXiv:1412.6980 [Cs]*, December. <http://arxiv.org/abs/1412.6980>.
- 559 Kolberg, Matthias, Kari R. Strand, Pål Graff, and K. Kristoffer Andersson. 2004. "Structure, Function,  
560 and Mechanism of Ribonucleotide Reductases." *Biochimica et Biophysica Acta - Proteins and*  
561 *Proteomics*. <https://doi.org/10.1016/j.bbapap.2004.02.007>.
- 562 Kuscu, Cem, Sevki Arslan, Ritambhara Singh, Jeremy Thorpe, and Mazhar Adli. 2014. "Genome-Wide  
563 Analysis Reveals Characteristics of off-Target Sites Bound by the Cas9 Endonuclease." *Nature*  
564 *Biotechnology* 32 (7): 677–83. <https://doi.org/10.1038/nbt.2916>.
- 565 Labuhn, Maurice, Felix F. Adams, Michelle Ng, Sabine Knoess, Axel Schambach, Emmanuelle M.  
566 Charpentier, Adrian Schwarzer, Juan L. Mateo, Jan-Henning Klusmann, and Dirk Heckl. 2018.  
567 "Refined SgRNA Efficacy Prediction Improves Large- and Small-Scale CRISPR–Cas9  
568 Applications." *Nucleic Acids Research* 46 (3): 1375–85. <https://doi.org/10.1093/nar/gkx1268>.
- 569 Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature*  
570 *Methods* 9 (4): 357–59. <https://doi.org/10.1038/nmeth.1923>.
- 571 Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo  
572 Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The



- 573 Sequence Alignment/Map Format and SAMtools." *Bioinformatics (Oxford, England)* 25 (16):  
574 2078–79. <https://doi.org/10.1093/bioinformatics/btp352>.
- 575 Lioy, Virginia S., Axel Cournac, Martial Marbouty, Stéphane Duigou, Julien Mozziconacci, Olivier  
576 Espéli, Frédéric Bocard, and Romain Koszul. 2018. "Multiscale Structuring of the E. Coli  
577 Chromosome by Nucleoid-Associated and Condensin Proteins." *Cell* 172 (4): 771-783.e18.  
578 <https://doi.org/10.1016/j.cell.2017.12.027>.
- 579 Moreno-Mateos, Miguel A., Charles E. Vejnár, Jean-Denis Beaudoin, Juan P. Fernández, Emily K. Mis,  
580 Mustafa K. Khokha, and Antonio J. Giraldez. 2015. "CRISPRscan: Designing Highly Efficient  
581 SgRNAs for CRISPR-Cas9 Targeting *in Vivo*." *Nature Methods* 12 (10): 982–88.  
582 <https://doi.org/10.1038/nmeth.3543>.
- 583 San Millán, Alvaro, Karl Heilbron, and R. Craig Maclean. 2013. "Positive Epistasis between Co-Infecting  
584 Plasmids Promotes Plasmid Survival in Bacterial Populations." *The ISME Journal* 8 (10).  
585 Nature Publishing Group: 601–12. <https://doi.org/10.1038/ismej.2013.182>.
- 586 Schmittgen, Thomas D., and Kenneth J. Livak. 2008. "Analyzing Real-Time PCR Data by the  
587 Comparative CT Method." *Nat. Protocols* 3 (6). Nature Publishing Group: 1101–8.
- 588 Sharan, Shyam K., Lynn C. Thomason, Sergey G. Kuznetsov, and Donald L. Court. 2009.  
589 "Recombineering: A Homologous Recombination-Based Method of Genetic Engineering."  
590 *Nature Protocols* 4 (2): 206–23. <https://doi.org/10.1038/nprot.2008.227>.
- 591 Sternberg, S. H., S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna. 2014. "DNA Interrogation by  
592 the CRISPR RNA-Guided Endonuclease Cas9." *Nature* 507 (7490): 62–67.  
593 <https://doi.org/10.1038/nature13011>.
- 594 Sternberg, Samuel H., Benjamin LaFrance, Matias Kaplan, and Jennifer A. Doudna. 2015.  
595 "Conformational Control of DNA Target Cleavage by CRISPR–Cas9." *Nature* 527 (7576): 110–  
596 13. <https://doi.org/10.1038/nature15544>.
- 597 St-Pierre, François, Lun Cui, David G. Priest, Drew Endy, Ian B. Dodd, and Keith E. Shearwin. 2013.  
598 "One-Step Cloning and Chromosomal Integration of DNA." *ACS Synthetic Biology* 2 (9): 537–  
599 41. <https://doi.org/10.1021/sb400021j>.
- 600 Szczelkun, M. D., M. S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, and  
601 R. Seidel. 2014. "Direct Observation of R-Loop Formation by Single RNA-Guided Cas9 and  
602 Cascade Effector Complexes." *Proc Natl Acad Sci U S A* 111 (27): 9798–9803.  
603 <https://doi.org/10.1073/pnas.1402597111>.
- 604 Vvedenskaya, Irina O., Yuanchao Zhang, Seth R. Goldman, Anna Valenti, Valeria Visone, Deanne M.  
605 Taylor, Richard H. Ebright, and Bryce E. Nickels. 2015. "Massively Systematic Transcript End  
606 Readout, 'MASTER': Transcription Start Site Selection, Transcriptional Slippage, and  
607 Transcript Yields." *Molecular Cell* 60 (6): 953–65.  
608 <https://doi.org/10.1016/j.molcel.2015.10.029>.
- 609 Wang, T., J. J. Wei, D. M. Sabatini, and E. S. Lander. 2014. "Genetic Screens in Human Cells Using the  
610 CRISPR-Cas9 System." *Science* 343 (6166): 80–84. <https://doi.org/10.1126/science.1246981>.
- 611 Wu, X., D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu, D. B. Dadon, A. W. Cheng, et al. 2014. "Genome-  
612 Wide Binding of the CRISPR Endonuclease Cas9 in Mammalian Cells." *Nat Biotechnol*, April.  
613 <https://doi.org/10.1038/nbt.2889>.
- 614 Xu, Han, Tengfei Xiao, Chen-Hao Chen, Wei Li, Cliff Meyer, Qiu Wu, Di Wu, et al. 2015. "Sequence  
615 Determinants of Improved CRISPR SgRNA Design." *Genome Research*, June, gr.191452.115.  
616 <https://doi.org/10.1101/gr.191452.115>.
- 617 Xu, T., Y. Li, Z. Shi, C. L. Hemme, Y. Li, Y. Zhu, J. D. Van Nostrand, Z. He, and J. Zhou. 2015. "Efficient  
618 Genome Editing in *Clostridium Cellulolyticum* via CRISPR-Cas9 Nickase." *Appl Environ*  
619 *Microbiol*, April. <https://doi.org/10.1128/AEM.00873-15>.
- 620