Published in final edited form as:
*Nature*. 2015 April 23; 520(7548): 505–510. doi:10.1038/nature14302.

# CRISPR adaptation biases explain preference for acquisition of foreign DNA

**Asaf Levy**[#1], **Moran G. Goren**[#2], **Ido Yosef**[2], **Oren Auster**[2], **Miriam Manor**[2], **Gil Amitai**[1], **Rotem Edgar**[2], **Udi Qimron**[2,†,#], and **Rotem Sorek**[1,†,#]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

[2]Department of Clinical Microbiology and Immunology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

[#] These authors contributed equally to this work.

## Abstract

In the process of CRISPR adaptation, short pieces of DNA ("spacers") are acquired from foreign elements and integrated into the CRISPR array. It so far remained a mystery how spacers are preferentially acquired from the foreign DNA while the self chromosome is avoided. Here we show that spacer acquisition is replication-dependent, and that DNA breaks formed at stalled replication forks promote spacer acquisition. Chromosomal hotspots of spacer acquisition were confined by *Chi* sites, which are sequence octamers highly enriched on the bacterial chromosome, suggesting that these sites limit spacer acquisition from self DNA. We further show that the avoidance of "self" is mediated by the RecBCD dsDNA break repair complex. Our results suggest that in *E. coli*, acquisition of new spacers depends on RecBCD-mediated processing of dsDNA breaks occurring primarily at replication forks, and that the preference for foreign DNA is achieved through the higher density of *Chi* sites on the self chromosome, in combination with the higher number of forks on the foreign DNA. This model explains the strong preference to acquire spacers from both high copy plasmids and phages.

CRISPR-Cas is an adaptive defense system in bacteria and archaea that provides acquired immunity against phages and plasmids [1-6]. It is comprised of multiple Cas genes, as well as

[#]Corresponding authors Correspondence and request for materials should be addressed to: Udi Qimron (ehudq@post.tau.ac.il) or Rotem Sorek (rotem.sorek@weizmann.ac.il) .
[†]These authors jointly supervised this work.

Contributions
M.G., U.Q., A.L. and R.S. conceived and designed the research studies; M.G., I.Y., O.A., M.M., G.A., and R.E. performed the experiments; A.L., M.G., U.Q., and R.S. analyzed data; A.L., M.G., U.Q., and R.S. wrote the manuscript.

Competing financial interests
The authors declare no competing financial interests.

Data deposition and accession
Raw data of spacer sequences are accessible through http://www.weizmann.ac.il/molgen/Sorek/files/CRISPR_adaptation_2015/crispr_adaptation_2015_data.html.
RNA-seq data are available in the NCBI SRA database under experiments accession numbers SRX862155-SRX862158 in study SRP053013.

an array of short sequences ("spacers") that are mostly derived from exogenous DNA and are interleaved by short DNA repeats. The CRISPR-Cas mode of action is divided into three main stages: Adaptation, Expression and Interference. In the adaptation stage, a new spacer is acquired from the foreign DNA and integrated into the CRISPR array. In the expression stage the repeat-spacer array is transcribed and further processed into short crRNAs. These mature crRNAs, in turn, bind to Cas proteins and form the effector protein-RNA complex. During the interference stage the effector complex identifies foreign nucleic acid via base pairing with the crRNA and targets it for degradation.

Numerous recent studies have characterized the molecular mechanisms governing the expression and interference stages of the CRISPR activity, but the molecular details of the primary adaptation stage are still elusive. It was shown that the Cas1 and Cas2 proteins are necessary for primary spacer acquisition [7], and that they form a single active complex [8]. Several systems to study spacer acquisition in the model bacterium *E. coli* have been established[7-13]. Some of these systems only express Cas1 and Cas2 but lack the CRISPR interference machinery, so that the protospacer-contributing DNA molecule is not targeted for degradation [7,8,11-13]. Strikingly, despite the lack of selection against spacer acquisition from the self chromosome, the vast majority of spacers acquired in such interference-free systems are derived from the plasmid [7,8,11], suggesting an intrinsic preference for the Cas1+2 complex to acquire spacers from the exogenous DNA. The mechanism by which the Cas1+2 complex preferentially recognizes the foreign DNA as a source for acquisition of new spacers, while avoiding taking spacers from the self chromosome, remains a major unresolved question.

## Preference for exogenous DNA

We set out to understand the mechanism governing the self/non-self discrimination of the DNA source for spacer acquisition during the adaptation stage. For this, we used a previously described experimental system that monitors spacer acquisition *in vivo* in the *E. coli* type I-E CRISPR system [7,12]. In this system, *cas1* and *cas2* are carried on a plasmid (pCas1+2) and their expression is regulated by an arabinose-inducible T7 RNA polymerase (Extended Data Fig. 1). We have previously shown that expression of Cas1+2 in this system leads to spacer acquisition, i.e., expansion of the chromosomally encoded CRISPR I array in *E. coli* BL21-AI [7]. Since this strain of *E. coli* harbors a CRISPR array but lacks any *cas* genes on its genome, this system is interference-free, and thus does not allow 'primed' CRISPR adaptation [9,10,14,15].

Following overnight growth of an *E. coli* BL21-AI culture carrying pCas1+2, we amplified the leader-proximal end of the CRISPR I array using a forward primer on the leader and a reverse primer matching spacer 2 of the native array. The amplification product, containing both native and expanded arrays, was sequenced using low coverage Illumina technology (MiSeq) to accurately quantify the fraction of arrays that acquired a new spacer in each experiment. In parallel, high coverage Illumina sequencing (HiSeq) was performed on gel-separated expanded arrays, in order to characterize the source, location, and frequency of newly acquired spacers in high resolution (Extended Data Fig. 1). Overall, over 38 million newly acquired spacers were sequenced in this study (Extended Data Tables 1-3).

In cultures overexpressing Cas1 and Cas2 for 16 hours, 36.92% (±1.2) of the sequenced arrays contained a new spacer. Conversely, in cultures where Cas1+2 were not induced, 2.61% (±0.5) of the arrays contained a new spacer after 16 hours of incubation, indicating that the leakage of Cas1+2 transcription (as measured by RNA-seq, Supplementary Table 1) still resulted in spacer acquisition in a significant fraction of the cells (Extended Data Table 1a). Examining the origin of new spacers showed strong preference for spacer acquisition from the plasmid, with only 22.86% (±0.46) and 1.8% (±0.03) of the spacers derived from the self chromosome in the induced and non-induced cultures, respectively (Extended Data Table 1b). Considering the size of the plasmid (4.7kb) and its estimated copy number of 20-40, this represents 100-1000 fold enrichment for acquisition of spacers from the plasmid, as compared to what is expected by the DNA content in the cell. These results also show that lower expression of Cas1+2 leads to higher specificity for exogenous DNA. Therefore, most of the analyses henceforth are based on spacers acquired in conditions in which Cas1+2 are expressed but not overexpressed.

## Replication-dependent adaptation

Although only a small minority of spacers was derived from the *E. coli* chromosome, the extensive number of sequenced spacers allowed us to examine chromosome-scale patterns of spacer acquisition. Remarkably, strong biases in spacer acquisition were observed, defining several protospacer hotspots (Fig. 1a). As the protospacer adjacent motif (PAM) density on the chromosome scale is largely uniform (Fig. 1b, Extended Data Fig. 2), these protospacer hotspots could not be explained by excessive localization of PAM sequences in specific areas of the genome. We further investigated each of the hotspots in search for a mechanism that would explain the observed biases.

Spacer acquisition was more pronounced at areas closer to the chromosomal origin of replication (oriC), with a clear gradient of reduced protospacer density as a function of the distance from oriC (Fig. 1a). In replicating cells the DNA next to the oriC is replicated first, and hence the culture inevitably contains more copies of the origin-proximal DNA [16]. Indeed, upon sequencing of total genomic DNA extracted from the *E. coli* BL21-AI culture, we observed a gradient in the DNA content reminiscent of the protospacer gradient (Extended Data Fig. 2). Therefore, this oriC-centered spacer acquisition bias can largely be expected based on the average DNA content in the culture and, accordingly, normalizing protospacer density to DNA content eliminated most of the oriC-centered protospacer gradient (Fig. 1b).

The most striking protospacer hotspot was observed around the chromosomal replication terminus (Ter), in two major peaks showing ~7-20 fold higher protospacer density than the surrounding area (Fig.1b-c). The Ter macrodomain is the area where the two replication forks coming from opposite directions on the chromosome meet, leading to chromosome decatenation [17]. This chromosomal macrodomain contains unidirectional fork stalling sites called Ter sites (primarily *TerA* and *TerC*), which, during replication, stall the early-arriving replication fork until the late fork arrives from the other side [17]. We found that the primary fork-stalling sites *TerA* and *TerC* were the exact boundaries of the spacer acquisition hotspots (Fig. 1c). Moreover, the protospacer hotspots next to Ter sites were asymmetric relative to the fork direction of progression, with strong protospacers enrichment observed upstream to each fork stalling site and a relatively low, background protospacer density downstream to the stalled fork (Fig. 1b-c). Engineering of a native *Ter* site into the *pheA* locus on the bacterial chromosome generated a new localized protospacer hotspot, strongly supporting that hotspots for spacer acquisition directly correlate with replication fork stalling sites (Fig. 1d).

The correlation between spacer acquisition biases and the replication fork stalling sites may suggest that CRISPR adaptation is promoted by active replication of the protospacer-containing DNA. We conducted a series of experiments to test this hypothesis. First, we used the replication-inhibitor quinolone nalidixic acid on *E. coli* BL21-AI cells during induction of Cas1+2. As a control, we applied the RNA polymerase inhibitor rifampicin that blocks transcription in *E. coli* but allows DNA replication (this antibiotic does not interfere with transcription of Cas1+2 by the T7 RNA polymerase). Application of nalidixic acid resulted in an almost complete elimination of spacer acquisition (164x-fold reduction), but only ~2-fold reduction in spacer acquisition rates was observed in the rifampicin-treated cells (Fig. 2a; Extended Data Table 1c), providing support to the hypothesis that spacer acquisition depends on DNA replication.

To further substantiate these observations, we examined the acquisition rates in *E. coli* K-12 cells carrying the temperature sensitive allele *dnaC2* [18]. In these cells, initiation of DNA replication is blocked at 39°C but is permitted at 30°C. These cells were transformed with a pBAD-Cas1+2 vector, in which the Cas1+2 operon is directly controlled by an arabinose-inducible promoter. Since these cells encode the full set of Cas genes, the *casC* gene was also knocked out to avoid CRISPR interference or priming. As a control, we used an isogenic K-12 strain encoding the WT *dnaC* gene. After overnight growth in the replication-permissive temperature the two strains showed similar rates of spacer acquisition. However, when the temperature sensitive *dnaC2* cells were grown at 39°C, acquisition was almost completely abolished, with less than 0.1% of the sequenced arrays found to be expanded (Fig. 2b; Extended Data Table 2a). These results further strengthen the hypothesis that Cas1+2-mediated spacer acquisition in the *E. coli* type I-E CRISPR system requires active replication of the protospacer-containing DNA.

We next asked whether spacer acquisition preferences correlate with the position of the replication fork. For this, we transferred a culture of the temperature sensitive *dnaC2* cells to 39°C for 70 minutes. Since in this temperature replication re-initiation is inhibited, after 70 minutes there are no more progressing forks in these cells. We then induced Cas1+2 expression for 30 minutes, and transferred the culture to 30°C, resulting in synchronized initiation of replication. At these conditions, it takes the replication forks on average ~60 minutes to complete a full DNA replication cycle in *dnaC2* cells [19]. In accordance, we sequenced the newly acquired spacers at 20, 40, 60, 90 and 120 minutes following synchronous replication initiation. Strikingly, the fraction of spacers derived from the Ter region has gradually increased with the progression of the replication cycle, reaching 31% after 60 minutes (compared to only 6.4% at the 20 minutes time point; Fig. 2c; Extended Data Fig. 3; Extended Data Table 2b). The pattern repeated itself in the second cycle of replication (90 and 120 minutes; Fig 2c). These results demonstrate temporal correlation between the predicted position of stalled replication forks and the preference to acquire spacers from that position.

Combined, the above results support a model where the Cas1+2 complex has preference for acquiring spacers from the area of a stalled replication fork during DNA replication. This model is intriguing, as it largely explains the observed preference for spacer acquisition from high copy number plasmids. During DNA replication in a cell, the chromosome occupies two replication forks traveling from the *oriC* to the Ter, where their stalling will promote spacer acquisition. At the same replication cycle, each copy of the plasmid will occupy a traveling fork, which will also be stalled during the termination of plasmid replication (in a Ter-independent manner[20]). Therefore, the vast majority of stalled forks in a replicating cell localize to the multiple plasmid copies, and, if spacer acquisition is promoted by fork-stalling, the probability to acquire spacers from the plasmid is much higher. The model is in line with previous observations in *Sulfolobus*, showing that spacer acquisition from an infective virus does not occur unless the viral DNA is being replicated [21].

## Involvement of the DNA repair machinery

Another hotspot for spacer acquisition was observed just upstream of the CRISPR I array in the *E. coli* BL21-AI genome (Fig. 3a). This CRISPR-associated protospacer hotspot clearly depends on the CRISPR activity, because no hotspot was observed near the *E. coli* BL21-AI CRISPR II array, which lacks a leader sequence and is hence inactive [7] (Fig. 3a). Indeed, in *E. coli* K-12, where both arrays are known to be active, spacer acquisition assays showed a protospacer peak upstream to each of the two arrays (Fig. 3b). The protospacer peaks at the CRISPR region resembled the peaks seen at the Ter sites, in the sense that they were asymmetric with respect to the replication fork direction, implying that activity at the CRISPR array forms a replication fork stalling site. Presumably the DNA nicking occurring after the leader during insertion of a new spacer [13], stalls the replication fork thus generating a fork-dependent hotspot for spacer acquisition. Frequent stalling of the fork at the CRISPR would mean that the fork coming from the other direction will often be stalled for a longer time at the respective Ter site, *TerC*, waiting for the fork coming from the CRISPR direction to arrive (Extended Data Fig. 4). This may be one of the factors explaining why the *TerC* site is a much more pronounced protospacer hotspot than the *TerA* site (Fig. 1b-c). Another factor that can contribute to the observed *TerC*/*TerA* bias may be that the clockwise replichore in *E. coli* (*oriC* to *TerA*) is longer than the counter clockwise one (*oriC* to *TerC*), leading the forks to naturally stall at *TerC* more often than at *TerA*.

All of the spacer acquisition hotspots described above were defined by distinct peaks of high protospacer density, with peak widths ranging between 10-50kb (Fig. 3). On one end, these peaks were bounded by a fork stalling site, but the mechanism defining the boundary at the other end of the peaks was not clear. Strikingly, when searching for sequence motifs that preferentially appear at the other end of the peaks we found that all protospacer peaks were immediately flanked by the 8-mer motif GCTGGTGG, which is the canonical sequence of the *Chi* site (Fig. 3a-d). *Chi* sites interact with the double-strand break repair helicase/ nuclease complex RecBCD and regulate the repair activity [22]. When a dsDNA break occurs, RecBCD localizes to the exposed end, and then unwinds and degrades the DNA until reaching a *Chi* site [23]. Upon recognition of the *Chi* site, RecBCD generally ceases to degrade the DNA, and instead yields a ssDNA that is bound by RecA and invades a homologous duplex DNA, which forms a template for completion of the missing DNA [23]. *Chi* sites work in an asymmetric manner, meaning that the GCTGGTGG motif will only interact with RecBCD coming from the right-end of the DNA molecule (downstream to the site), whereas the reverse complement of *Chi* will only interact with RecBCD complexes coming from the left-end of the DNA [22]. RecBCD indiscriminately degrades linear DNA, including phage DNA, and it was therefore suggested that this complex is one of the lines of defense against phages [23]. Since *Chi* sites occur every ~5kb in the *E. coli* genome, which is ~14 times more frequent than expected by chance, these sites were suggested as markers of bacterial “self”, preventing RecBCD from excessively degrading the chromosome following dsDNA breaks [23].

Our results show that protospacer hotspots are defined between sites of stalled forks and *Chi* sites (Fig. 3). Stalled replication forks are known to be major hotspots for dsDNA breaks [24,25], and it was demonstrated that the vast majority of dsDNA breaks in bacteria occur during DNA replication [23, 26]. These data therefore may imply that Cas1+2 acquires spacers from degradation intermediates of RecBCD activity during the processing of dsDNA breaks that frequently occur at stalled replication forks.

Several lines of evidence support this hypothesis. First, the orientation of the *Chi* sites at the protospacer peaks was always consistent with the dsDNA break occurring at the fork direction rather than the other side, and the first properly oriented *Chi* site upstream to the stalled fork was always the site of peak boundary (Fig. 3a-d). Second, even outside the strong protospacer hotspots, there was a significant asymmetry in protospacer density upstream and downstream *Chi* sites (Fig. 4a). The effect of this asymmetry was seen up to a distance of about 5-10kb from the *Chi* site, consistent with an average distance of ~5kb between *Chi* sites in the *E. coli* genome [22]. Third, inducing a single, site specific dsDNA break in the chromosome using the homing endonuclease I-*Sce*I resulted in a clear protospacer hotspot that peaked at the site of the dsDNA break and was confined by *Chi* sites in the proper orientations (Fig. 4b), directly linking dsDNA breaks to spacer acquisition hotspots. Fourth, co-immunoprecipitation assays suggested that Cas1 interacts with RecB and RecC 27 (although these interactions were not verified using purified proteins), supporting a model where the Cas1+2 complex is directly fed from RecBCD DNA degradation products. And finally, Cas1 was shown to efficiently bind ssDNA, which is amply generated during RecBCD DNA processing activity [23,27].

To test whether spacer acquisition indeed depends on the activity of the RecBCD complex, we used *E. coli* strains in which *recB*, *recC* or *recD* were deleted. Deep-sequencing-based quantification of spacer acquisition rates in these mutants showed reduced acquisition in all of these deletion strains (Fig. 4e; Extended Data Table 3a). Moreover, analysis of chromosomal protospacers in these mutants showed loss of spacer acquisition asymmetry near *Chi* sites (Fig. 4c), resulting in broader protospacer hotspots on the self chromosome (Fig. 4d). In accordance, the fraction of spacers derived from the self chromosome was ~10-fold higher in the *recB*, *recC* and *recD* deletion strains as compared to the WT strain (Fig.

4f; Extended Data Table 3a). These results show that CRISPR adaptation is partially dependent on the activity of the RecBCD dsDNA break repair complex, and that this activity is also responsible for some of the self/non-self discrimination properties of the CRISPR adaptation process. Consistent with these results, expression of a RecBCD inhibitor protein, the product of gene 5.9 of the T7 bacteriophage 28, showed reduced spacer acquisition as compared to exogenous expression of a control protein (Extended Data Fig. 5).

It is noteworthy that in *recB* and *recC* deletions, the RecBCD complex is entirely nonfunctional, whereas the *recD* deletion produces a complex, RecBC, that is fully functional for DNA unwinding but entirely lacks nuclease activity[23]. Our observation that the *recD* deletion mutant has poor spacer acquisition activity suggests that the nuclease activity of the RecBCD enzyme is important for spacer acquisition and implies that the degradation products generated by RecBCD during DNA processing between dsDNA break and a *Chi* site may be the source for new spacers.

The involvement of *Chi* sites, as points where spacer acquisition activity is terminated, provides another axis for the avoidance of self DNA in CRISPR adaptation. Since the pCas plasmid is completely devoid of *Chi* sites, its DNA will be fully degraded by RecBCD following any dsDNA break, providing plenty of potential substrate for Cas1+2. In contrast, the high density of *Chi* sites on the bacterial chromosome serves for the relative avoidance of Cas1+2 to acquire spacers from the chromosome, because RecBCD will only degrade the chromosomal DNA until reaching the nearest *Chi* site (Fig. 5a-b). Indeed, the ~10 fold higher acquisition frequency from the self choromosome seen in the *recB*, *recC* and *recD* deletion strains conforms with the natural 14-fold enrichment of *Chi* sites on the chromosome. To further examine whether *Chi* sites limit spacer acquisition, we performed spacer acquisition experiments with a plasmid that was engineered to contain a cluster of 4 consecutive *Chi* sites. As expected, an increased preference for chrmomosomal DNA in spacer acquisition was measured for the *Chi*-containing plasmid (Fig 4g; Extended Data Table 3b; Extended Data Fig. 6).

In conclusion, these results converge to a single, unifying model that explains the preference of the CRISPR adaptation machinery to acquire spacers from foreign DNA, as well as the observed biases in spacer acquisition patterns (Fig. 5). Under this model, Cas1+2 takes the DNA substrate for spacer acquisition from degradation products of RecBCD activity during the processing of dsDNA breaks. Since the vast majority of dsDNA breaks in the cell occur during DNA replication [26] with stalled replication forks being major hotspots for such breaks [24,25], high copy number plasmids are much more prone to spacer acquisition due to the higher number of forks on plasmids (Fig. 5c). The high density presence of *Chi* sites on the bacterial chromosome further protects it from extensive spacer acquisition (Fig. 5b). Moreover, as most phages enter the cell as a linear DNA, and since RecBCD would bind any exposed linear DNA and process it until the nearest *Chi* site [22], unprotected phage DNA will be a target for spacer acquisition immediately upon entry to the cell, providing an additional preference for spacer acquisition specifically from phage DNA (Fig. 5d). If entry to the cell was successful the extensive replication activity of the phage DNA would provide another anchor for spacer acquisition from phage.

## Online Methods

### Reagents, strains, and plasmids

Luria-Bertani (LB) medium (10 g/l tryptone, 5 g/l yeast extract, and 5 g/l NaCl) and agar were from Acumedia. Antibiotics and L-arabinose were from Calbiochem. Isopropyl-beta-D-thiogalactopyranoside (IPTG) was from Bio-Lab. Calcium chloride ($CaCl_2$), sodium citrate (Na-citrate), restriction enzymes, T4 Polynucleotide Kinase (PNK) and Phusion high fidelity DNA polymerase were from New England Biolabs. Rapid ligation kit was from Roche. Taq DNA polymerase was from LAMDA biotech. NucleoSpin Gel and PCR Clean-up kit was from Macherey-Nagel (MN). The bacterial strains, plasmids, and oligonucleotides used in this study are listed in Supplementary Table 2.

### Plasmid construction

Plasmids were constructed using standard molecular biology techniques. DNA segments were amplified by PCR. Standard digestion of the PCR products and vector by restriction enzymes was carried out according to the manufacturer's instructions.

pBAD plasmid encoding Cas1 and Cas2 was constructed by amplifying Cas1 and Cas2 from pWUR399 plasmid [29] using oligonucleotides IY86F and IY86R (Supplementary Table 2). The amplified DNA and pBAD18 vector were both digested by *Sac*I and *Sal*I and ligated to yield pBAD-Cas1+2. The DNA insert was sequenced to exclude mutations introduced during cloning. pWUR plasmid encoding Cas1 and Cas2 under lac promoter was constructed by amplifying the lac promoter from pCA24N plasmid [29] using oligonucleotides SM18F and OA11R and amplifying the pCas1+2 vector using oligonucleotides IY56F and OA12F (Supplementary Table 2). The amplified products were ligated to yield pCas1+2-IPTG and sequenced to exclude mutations introduced during cloning. pWURV2 plasmid was constructed by amplifying the pCas1+2 backbone [29] using oligonucleotides IY81F and IY56R (Supplementary Table 2) followed by self-ligation. pCas1+2 plasmids harbouring 4 *Chi* sites/non *Chi* sites were constructed by annealing the oligonucleotide MM1F to MM1R or MM2F to MM2R, respectively, and ligating the dsDNA product to *Nco*I digested pCas1+2. pBAD33-gp5.9 plasmid encoding the T7 gene 5.9 was constructed by amplifying the 5.9 gene from the T7 bacteriophage using oligonucleotides RE45F and IY256R (Supplementary Table 2). The amplified DNA and pBAD33 vector were both digested by *Sca*I and *Sal*I and ligated to yield pBAD33-gp5.9.

### Strain construction using recombination-based genetic engineering (recombineering)

BL21-AI *recB/C/D* deletion mutants were constructed using recombineering method, as described previously [37]. Briefly, an overnight culture of BL21-AI/pSIM6 [38] was diluted 75-fold in 250 ml LB + 100 μg/ml ampicillin and aerated at 32 °C. When the $OD_{600}$ reached 0.5, the culture was heat-induced for recombination function of the prophage at 42 °C for 15 min in a shaking water bath. The induced culture was immediately cooled on ice slurry and then pelleted at 4600 x *g* at 0 °C for 10 min. The pellet was washed three times in ice-cold $ddH_2O$, then resuspended in 200 μl ice-cold $ddH_2O$ and kept on ice until electroporation with ~300 ng of a gel-purified PCR product encoding the construct specified in Supplementary Table 2 containing a kanamycin-resistance cassette flanked by 50 bp homologous to the desired insertion site. A 50 μl aliquot of electrocompetent cells was used for each electroporation in a 0.2-cm cuvette at 25 μF, 2.5 kV and 200 Ω. After electroporation, the bacteria were recovered in 1 ml of 2YT (16 g/l bacto-tryptone, 10 g/l yeast extract, 5 g/l NaCl) for 2 h in a 32 °C shaking water bath and inoculated on selection plates containing 25 μg/ml kanamycin. The DNA insertion into the resulting strains was confirmed by DNA sequencing of a PCR product amplifying the region.

### Transductions

P1 Transductions were used for replacing *araB* with a cassette encoding the T7-RNA polymerase linked to tetracycline resistance marker, or *thr* with *dnaC2* allele linked to *Tn10* encoding tetracycline resistance marker, or *pheA* with TerB site linked to Spectinomycin. P1 lysate was prepared as followed: overnight cultures of donor strain BL21-AI (for T7 RNA polymerase) or MG1655dnaC2 (for *dnaC2* allele) [31] or JJC1819 (for *pheA*::TerB-Spec) were diluted 1:100 in 2.5 ml LB + 5 mM $CaCl_2$. After 1 h shaking at 37 °C (or 30 °C for MG1655dnaC2), 0 to 100 μl phage P1 was added. Cultures were aerated for 1 to 3 h, until lysis occurred. The obtained P1 lysate was used in transduction where 100 μl fresh overnight recipient culture was mixed with 1.25 μl of 1 M $CaCl_2$ and 0 to 100 μl P1 phage lysate. After incubation for 30 min at 30 °C without shaking, 100 μl Na-citrate and 500 μl LB were added. Cultures were incubated at 37 °C or 30 °C for 45 or 60 min, respectively, then 3 ml of warm LB supplemented with 0.7% agar was added and the suspension was poured onto a plate containing the appropriate drug. Transductants obtained on antibiotic plates were streaked several times on selection plates and verified by PCR for the presence of the transduced DNA fragment.

### Markerless insertion of I-*Sce*I restriction site into the genome

A linear DNA containing the *Kan-sacB* cassette [39] for kanamycin resistance and sucrose sensitivity was amplified by PCR with oligonucleotides MG53F and MG53R that provided homology to a region downstream to the *ydhQ* gene. The *Kan-sacB* cassette was inserted into DY378 strain [40] by recombineering (as described above). Colonies that were found to be resistant to kanamycin and sensitive to sucrose, i.e., containing the *Kan-sacB* cassette were picked and verified by PCR. The *Kan-sacB* cassette was transferred by P1 transduction from DY378 to BL21-AI. A second PCR was performed using oligonucleotides MG54F and MG54R that produce a short linear DNA containing the I-*Sce*I restriction site with homology of 50 bp upstream and downstream to the *ydhQ* stop codon. Recombineering of this DNA fragment to BL21-AI, *ydhQ-Kan-sacB* resulted in kanamycin sensitive and sucrose resistant colonies that replaced the *Kan-sacB* cassette with I-*Sce*I restriction site immediately after the *ydhQ* stop codon. DNA from the resulting strain was sequence-verified for the presence of an intact I-*Sce*I site.

### CRISPR array size determination prior to acquisition assay

All strains underwent a preliminary validation step aimed to rule out acquisition prior to induction: *E. coli* BL21-AI or K-12 harboring pCas1+2 or pBAD-Cas1+2 plasmids, respectively, were spread on LB + 50 μg/ml streptomycin or 100 μg/ml ampicillin + 0.2% (wt/vol) glucose plates and incubated overnight at 37 °C or 30 °C (for K12Δ*casCdnaC2*). A single colony was picked from each plate and used as template in a PCR amplifying CRISPR array I for BL21-AI or array II for K-12. Primers MG7R/OA1R and MG7R/MG34F were used to detect array expansion for BL21-AI and K-12, respectively (Supplementary Table 2). Only colonies that did not undergo array expansion were used in the acquisition assays described below.

## Standard acquisition assay

A single colony of BL21-AI or BL21-AIΔ*pheA::terB* or BL21-AIΔ*recB/C/D* strains harboring pCas1+2 plasmid or BL21-AI strain harboring pWURV2 plasmid or K-12Δ*casC* T7RNAP strain harboring pBAD-Cas1+2 plasmid or BL21-AI *ydhQ*-I-*Sce*I site strain harboring pCas1+2-IPTG and pBAD-I-*Sce*I plasmids or BL21-AI strain harboring p*Chi* or pCtrl-*Chi* plasmids and BL21-AI strain harboring pCas1+2 and pBAD33-gp5.9 plasmids were inoculated in LB medium containing 50 μg/ml streptomycin + 0.2% (wt/vol) glucose for BL21-AI strains carrying a single plasmid or 100 μg/ml ampicillin + 0.2% (wt/vol) glucose for K12 strain or 100 μg/ml ampicillin + 50 μg/ml streptomycin + 0.2% (wt/vol) glucose for BL21-AI ydhQ-I-*Sce*I site/pCas1+2-IPTG/pBAD-I-*Sce*I strain or 200 μg/ml ampicillin + 35 μg/ml chloramphenicol for BL21-AI/ pCas1+2/ pBAD33-gp5.9. Cultures were aerated at 37 °C for 16 h. Each overnight culture was diluted 1:600 in LB medium containing appropriate antibiotics with or without 0.2% (wt/vol) L-arabinose + 0.1 mM IPTG for pCas1+2, p*Chi* and pCtrl-*Chi* harboring strains or 0.2% (wt/vol) L-arabinose for pBAD-Cas1+2 harboring strains or 0.02 mM IPTG for and 0% L-arabinose for pCas1+2-IPTG and pBAD-I-*Sce*I harboring strain or 0.4% (wt/vol) L-arabinose for pCas1+2 and pBAD33-gp5.9 harboring strain. Cultures were aerated at 37 °C for additional 16 h. DNA from these cultures was used as template (see - *DNA preparation for PCR*) in PCRs using primers OA1F/IY130R (PCR1) and RE10RD/IY230R (PCR2) for amplifying BL21-AI CRISPR array I or MG116F/MG34F (PCR1, see below) and RE10RD/MG115R (PCR2, see below) for amplifying K-12 array II.

## Acquisition assay in the presence of antibiotics

Single colony of BL21-AI/pCas1+2 was inoculated in LB medium containing 50 μg/ml streptomycin + 0.2% (wt/vol) glucose and aerated at 37 °C for 16 h. The overnight cultures were diluted 1:600 in LB medium containing 50 μg/ml streptomycin with or without 0.2% (wt/vol) L-arabinose + 0.1 mM IPTG and aerated at 37 °C. Once cultures reached $OD_{600}$ of 0.25, cells were centrifuged in a microcentrifuge for 10 min at 13,000 x *g* and resuspended in LB medium containing 50 μg/ml streptomycin or 50 μg/ml nalidixic acid or 100 μg/ml rifampicin with or without 0.2% (wt/vol) L-arabinose + 0.1 mM IPTG. Cultures were aerated for 16 h at 37 °C, lysed and served as template for PCR1 using primers OA1F/IY130R (PCR1) and RE10RD/IY230R (PCR2) for amplifying BL21-AI CRISPR array I.

## Acquisition assay in replication-deficient strains

Single colony of K-12Δ*casC* (control) or K-12Δ*casCdnaC2* harboring pBAD-Cas1+2 was inoculated in LB medium containing 100 μg/ml ampicillin + 0.2% (wt/vol) glucose and aerated at 30 °C, for 16 h. The overnight cultures were diluted 1:600 in LB medium containing 100 μg/ml ampicillin + 0.2% (wt/vol) L-arabinose and aerated at 30 °C or 39 °C for another 16 h. Cultures were then lysed and used as template in PCRs using primers MG116F/MG34F (PCR1) and RE10RD/MG115R (PCR2) for amplifying K-12 array.

### Synchronized acquisition assay

Single colony of K-12Δ*casC* (control) or K-12Δ*casCdnaC2* harboring pBAD-Cas1+2 was inoculated in LB medium containing 100 μg/ml ampicillin + 0.2% (wt/vol) glucose and aerated at 30 °C, for 16 h. The overnight cultures were diluted 1:600 in LB medium containing 100 μg/ml ampicillin + 0.2% (wt/vol) glucose and aerated at 30 °C until $OD_{600}$ reached 0.25. Cultures were then split into six tubes and transferred to non-permissive temperature − 39 °C. After 70 min, induction of Cas1+2 was performed: cells were centrifuged in a standard centrifuge (4,600 x *g*, 10 min), resuspended in LB medium containing 100 μg/ml ampicillin + 0.2% (wt/vol) L-arabinose and aerated for additional 30 min at 39 °C. Replication was then initiated by aerating the split cultures at 30 °C for – 0, 20, 40, 60, 90, 120 min. For replication arrest, cells were lysed and used as template in PCRs using primers MG116F/MG34F (PCR1) and RE10RD/MG115R (PCR2) for amplifying K-12 array.

### DNA preparation for PCR

DNA was prepared from all cultures that underwent acquisition assays. 1 ml of each culture was centrifuged in a microcentrifuge for 1 min at 13,000 x *g* and re-suspended in 100 μl LB medium. The concentrated culture underwent fast freeze in liquid nitrogen, boiled at 95 °C for 10 min and placed on ice for 5 min. The lysate was then centrifuged in a microcentrifuge for 2 min at 13,000 x *g*, the supernatant was transferred to a new tube and served as template for PCR1 (see *Preparation of DNA samples for deep sequencing*).

### Cultures preparation for RNA-seq

Single colony of *E. coli* BL21-AI strain harboring pCas1+2 plasmid was inoculated in LB medium containing 50 μg/ml streptomycin + 0.2% (wt/vol) glucose and aerated at 37 °C for 16 h. Each overnight culture was diluted 1:600 in LB medium containing appropriated antibiotics with or without: 0.2% (wt/vol) L-arabinose + 0.1 mM IPTG. Following overnight growth cultures 15 ml from each culture was centrifuged in a standard centrifuge (4,600 x *g*, 10 min), the supernatant was discarded and the pellet underwent fast freeze in liquid nitrogen. Cell pellets were then thawed and incubated at 37°C with 300 μl 2 mg/ml lysozyme (Sigma-Aldrich cat# L6876-1G) in Tris 10mM EDTA 1mM pH 8.0, and total nucleotides were extracted using the Tri-Reagent® protocol, according to manufacturers instructions (Molecular Research Center, Inc. cat# TR118). TURBO DNA-free™ Kit was used to eliminate DNA from the sample, according to the manufacturer instructions (Life technologies – Ambion cat# AM1907). Enrichment for mRNA was accomplished by using the Ribo-Zero™ rRNA Removal Kits (Illumina-Epicentre cat#MRZB12424). The enriched mRNA sample was then further purified using Agencourt® AMPure® XP magnetic beads (Beckman Coulter cat# A63881). Purified bacterial mRNA was then used as the starting material for the preparation of cDNA libraries for next-generation sequencing using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (NEB cat# E7420S).

The NEBNext® multiplex oligos for Illumina® Index primer set1 (NEB cat#E7335S) was used as the adapters for the library.

## Total DNA purification

Overnight cultures of *E. coli* BL21-AI or K-12Δ*casC* T7RNAP harbouring pCas1+2 or pBAD-Cas1+2 plasmid, respectively, were diluted 1:600 and aerated for 16h at 37 °C in LB medium containing 50 μg/ml streptomycin or 100 ampicillin μg/ml + 0.2% (wt/vol) glucose. These overnight cultures were then diluted 1:600 in LB medium containing 50 μg/ml streptomycin or 100 ampicillin μg/ml with 0.2% (wt/vol) L-arabinose + 0.1 mM IPTG or without inducers and aerated at 37 °C. Once cultures reached $OD_{600}$ of 0.5–0.6 3 ml were removed and used for total DNA purification using the Macherey-Nagel NucleoSpin Tissue kit. Total DNA samples were used for deep sequencing (MiSeq).

## Preparation of spacer PCR products for deep sequencing

DNA from bacterial cultures that underwent various acquisition assays was amplified in two consecutive PCRs termed PCR1 and PCR2. In PCR1, The reaction contained 20 μl of Taq 2× Master Mix master mix, 1 μl of 10 μM forward and reverse primers (see Supplementary Table 2), 4 μL of bacterial lysate, and 14 μL of double-distilled water. The PCR started with 3 min at 95 °C followed by 35 cycles of 20 s at 95 °C, 20 s at 55 °C, and 20 s at 72 °C. The final extension step at 72 °C was carried out for 5 min. Half of the PCR1 content (20 μl) was purified using the DNA clean-up kit and were used for standard library prep procedures followed by deep sequenced (MiSeq), while the other half (20 μl) was loaded on a 2% (wt/vol) agarose gel and electrophoresed for 60 min at 120 V. Following gel separation, the expanded band was excised from the gel and purified using the DNA clean-up kit. 1 ng from the extracted band served as a template for the PCR2 reaction aimed to amplify the expanded CRISPR array products. PCR2 contained 10 μl of Taq 2× Master Mix master mix, 0.5 μl of 10 μM forward and reverse primers (Supplementary Table 2), 1 ng of the gel-extracted DNA from PCR1, and double-distilled water up to 20 μl. PCR2 program was identical to that of PCR1. The entire PCR2 content was loaded on a 2% (wt/vol) agarose gel, electrophoresed, excised and purified from the gel using the same conditions as in PCR1.

## Detection of protospacer identity and acquisition level

The PCR products described above were used for preparation of Illumina sequencing libraries and were sequenced using HiSeq or MiSeq machines according to manufacturer instructions. Several samples were multiplexed together in the same sequencing run. Demultiplexing was done to the different samples based on the different Illumina barcodes and based on 3 bp barcode that was part of the original PCR primer.

Reads were mapped against the *E. coli* genome and pCAS plasmid using blastn (with parameters: −e 0.0001 −F F). For strain K-12 the Refseq accession NC_000913.2 was used and for strain BL21-AI (for which genomic sequence is unavailable) the *E. coli* BL21-Gold(DE3)pLysS AG was used (Refseq accession NC_012947.1).

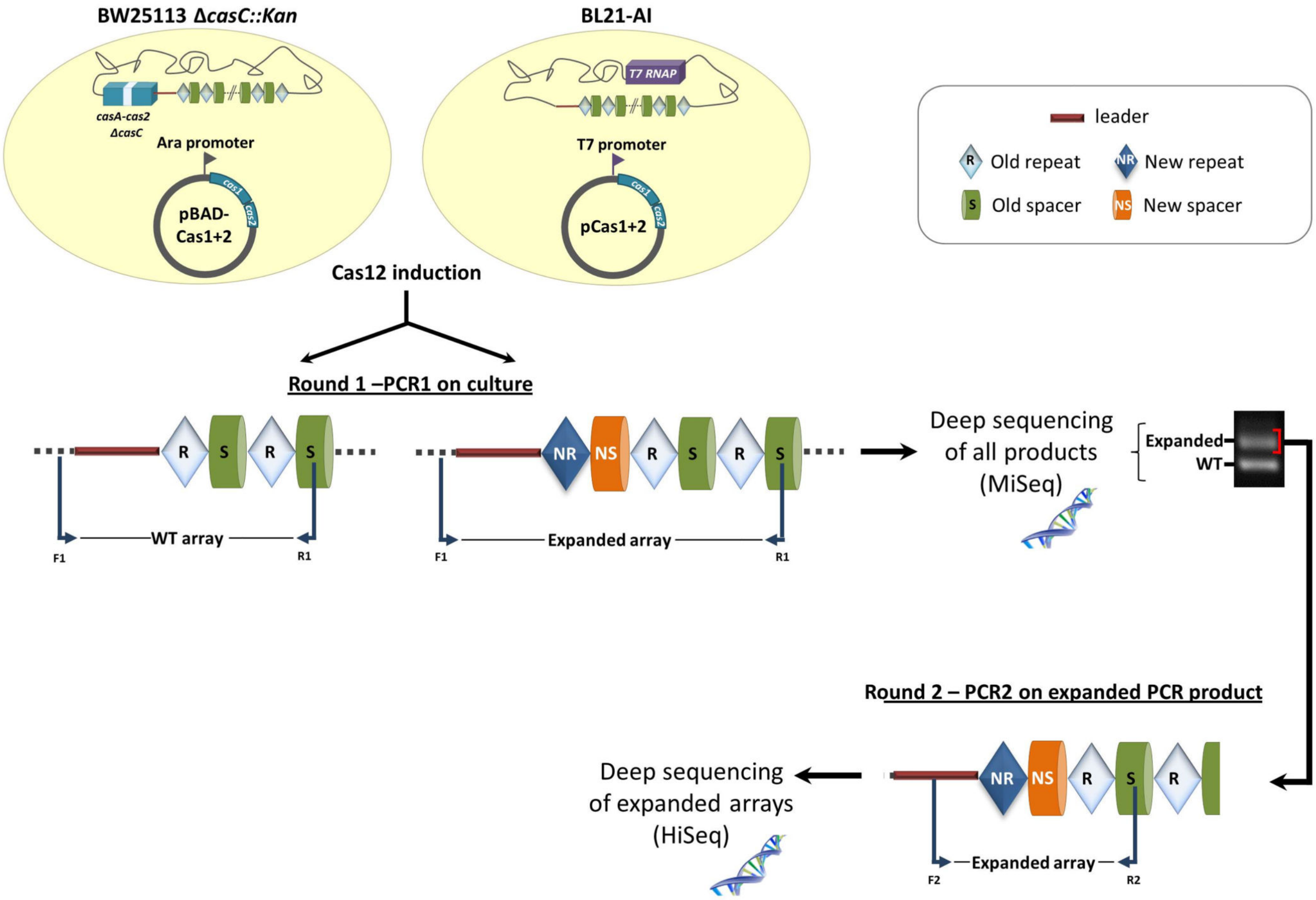
New spacer insertions were called based on sequence alignments of the resulting reads. For round 1 of the PCR (Extended Data Fig. 1) alignments supporting non-acquisition events were also recorded to quantify acquisition level. If the sequence read was fully mapped to the parental CRISPR locus in the leader-proximal side, a non-acquisition event was inferred. New acquisition events were inferred if the read alignment begun by a substring that was mapped to the CRISPR locus (‘pre-acquisition’ mapping) followed by a spacer-length substring that mapped elsewhere on the genome or the plasmid. Uninformative alignments, resulting from sequencing of the leader-distal side of the PCR amplicon were discarded. Spacer acquisition level for a sample was defined as the number of reads supporting acquisition events divided by the number of reads either supporting or rejecting spacer acquisition.

For round 2 of the PCR (enriching for expanded arrays only, Extended Data Fig. 1) we used only unambiguously mapped protospacers (e.g., spacers mapped to repetitive rRNA genes were discarded). In case that a spacer was mapped equally well both to the genome and the pCAS plasmid, only the plasmid protospacer position was used.

For the plots of protospacer distribution and hotspots (except for the plot in Extended Data Fig. 3), protospacer positions were recorded only once (meaning that if there were multiple spacers hitting the exact same position, the position was considered only once). This procedure was done in order to avoid biases stemming from PCR amplification of the CRISPR array, as well as local biases stemming from differential PAM preferences [12].

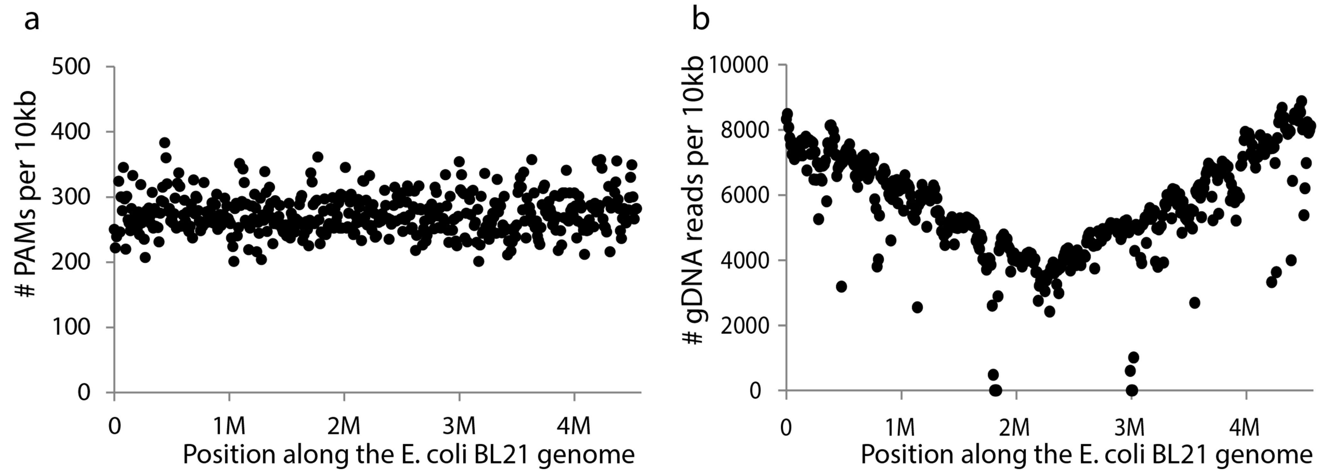
Data analysis was done using Perl and R scripts. Data visualization and statistical analysis was done using Excel and R, including the R circular package (http://cran.rproject.org/web/packages/circular/circular.pdf) for Figures 1 and Extended Data Fig. 4.

## Extended Data



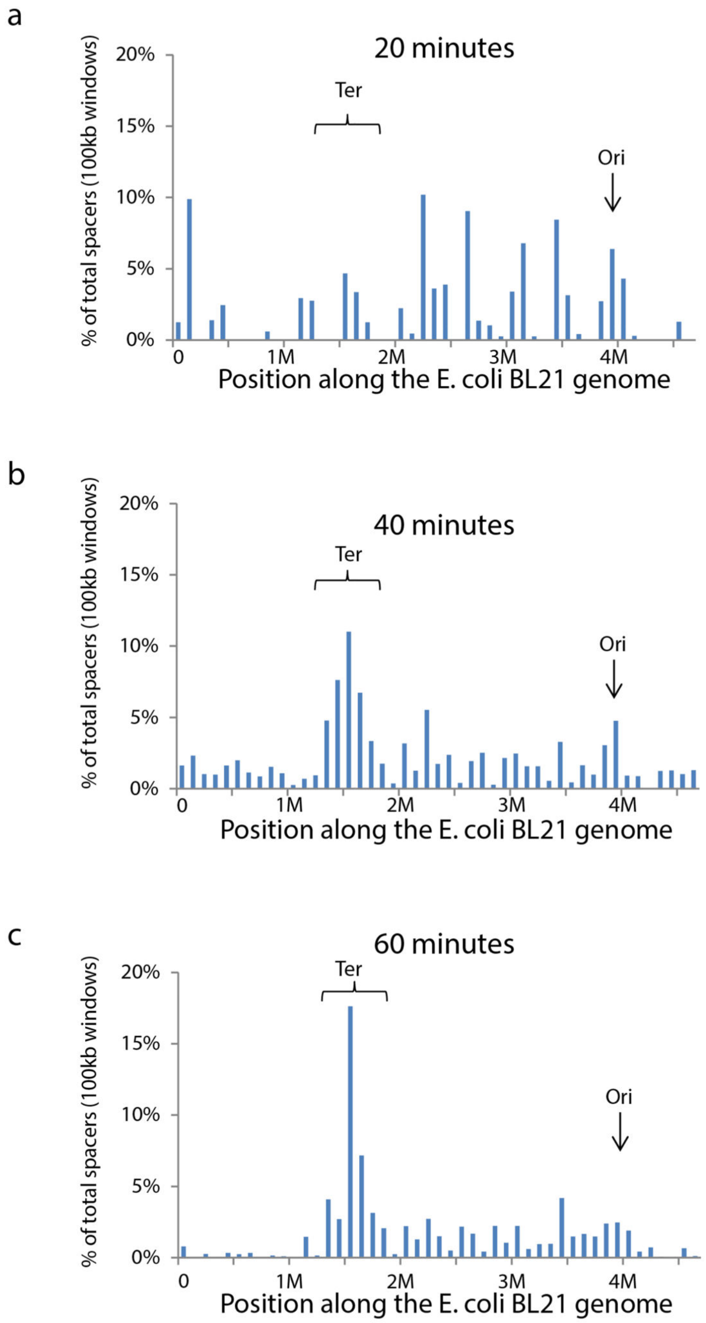
**Extended Data Figure 1. Graphic overview of the procedure for characterizing the frequency and sequence of newly acquired spacers**

DNA from cultures of either *E. coli* K-12 (left) or *E. coli* BL21-AI (right) strains expressing Cas1+2 from two different plasmids were used as templates for PCR. Round 1 was used to determine the frequency of spacer acquisition by comparing occurrences of expanded arrays to WT arrays. Round 2 amplified only the expanded arrays and, followed by deep sequencing, was used to determine the sequence, location, and source of newly acquired spacers.



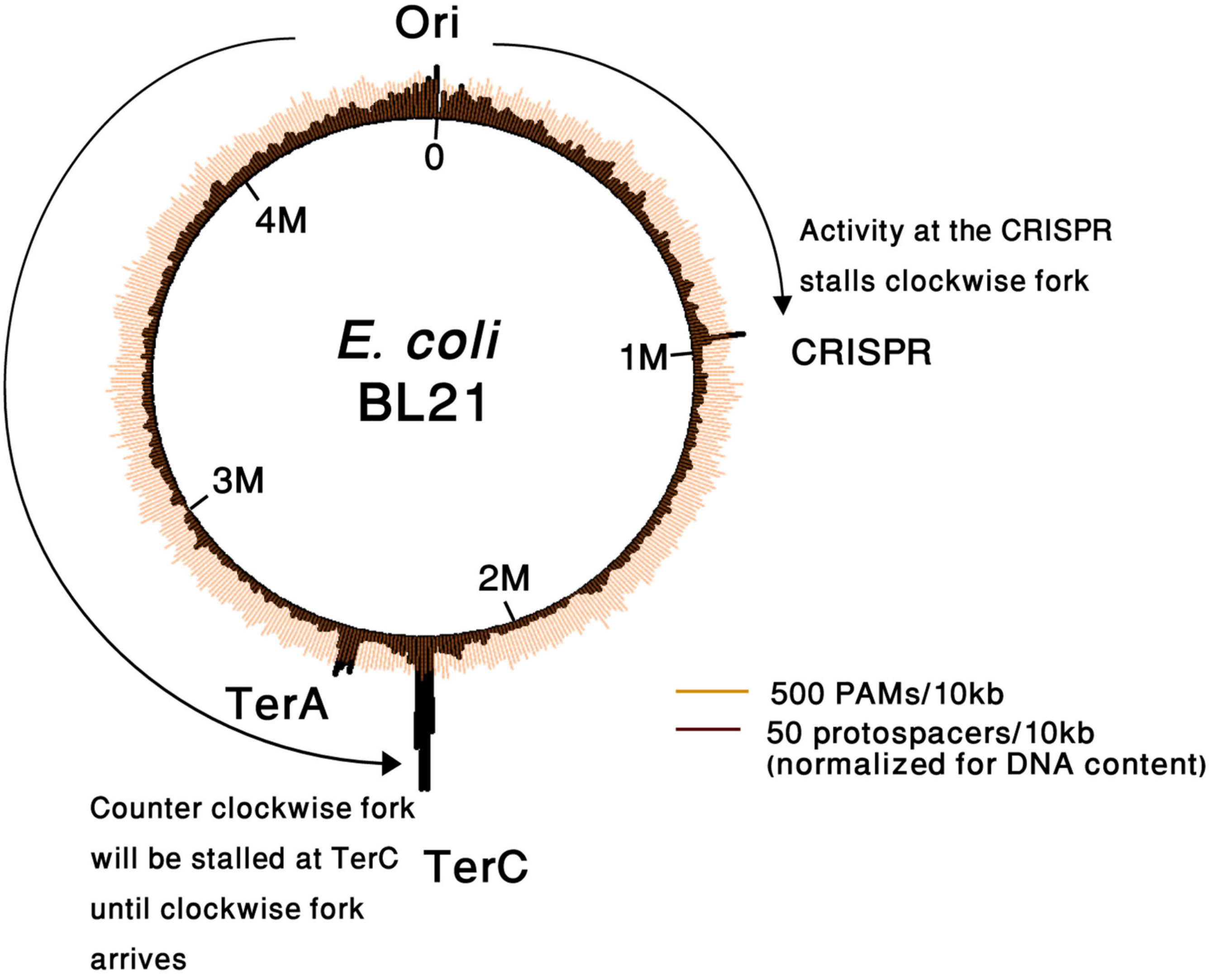
**Extended Data Figure 2. PAMs and DNA content along the *E. coli* BL21-AI genome**

(A) Distribution of PAM (AAG) sequences. Each data point represents the number of PAMs in a window of 10kb. (B) DNA content of a culture growing in log phase. Genomic DNA was extracted from *E. coli* BL21-AI cells carrying the pCas plasmid, grown at log phase, and was sequenced using the Illumina technology. The resulting reads were mapped to the sequenced *E. coli* BL21(DE3) genome (genbank accession NC_012947). Areas where little or no reads map to the genome represent regions that are present in the reference BL21(DE3) genome but are missing from the genome of the sequenced strain (BL21-AI).



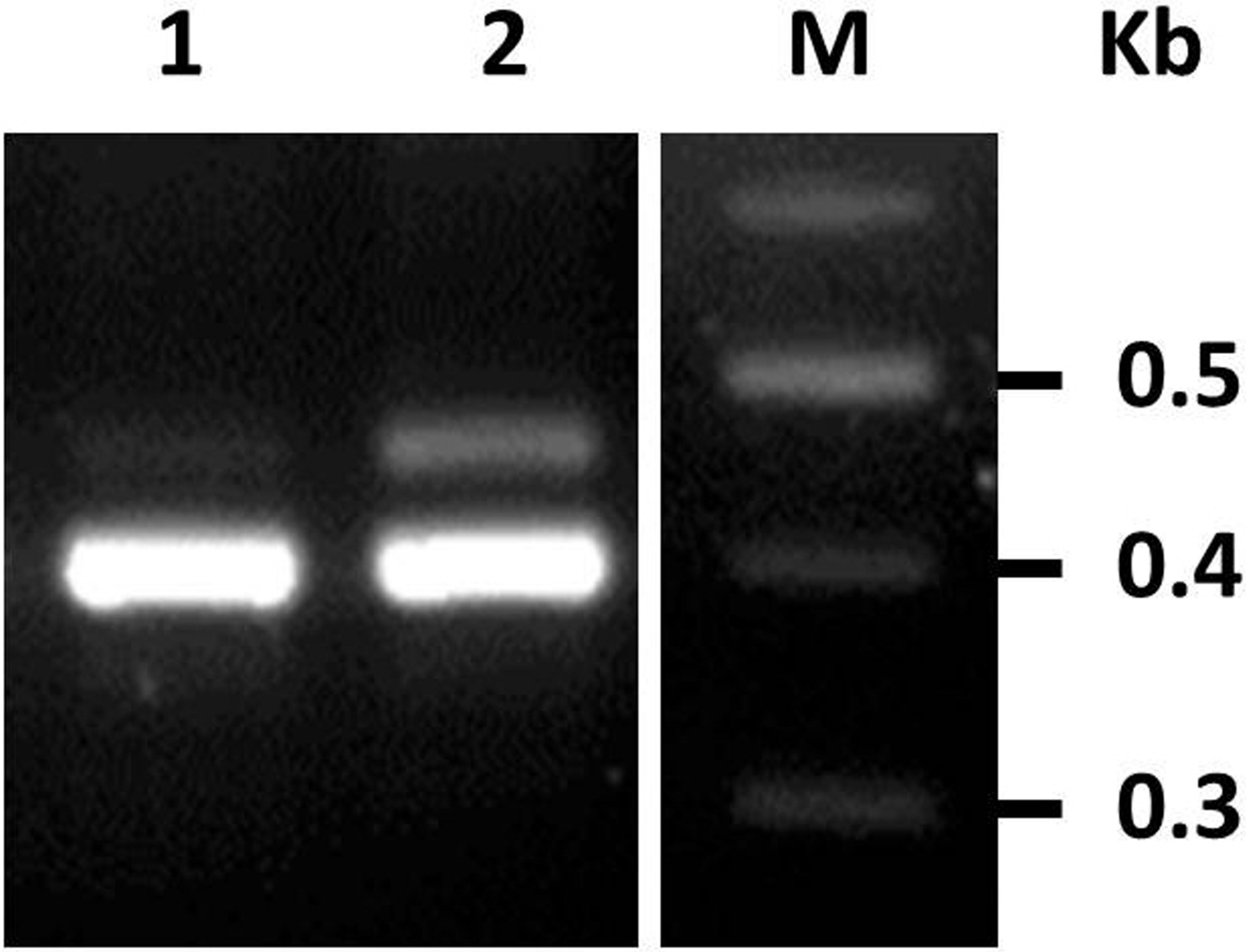
**Extended Data Figure 3. Distribution of newly acquired spacers on the genome during synchronized replication**

*E. coli* K-12Δ*casCdnaC2* cells were transferred from 39°C (replication restrictive temperature) to 30°C (replication permissive). Cas1+2 were induced in these cells 30 minutes prior to the transfer to 30°C and during the growth in 30°C. At given time points: (A) following 20 minutes; (B) following 40 minutes; (C) following 60 minutes from replication initiation, newly acquired spacers were sequenced. Shown are the positions of the newly acquired spacers in windows of 100kb, and their fraction out of the total new spacers in the sample.



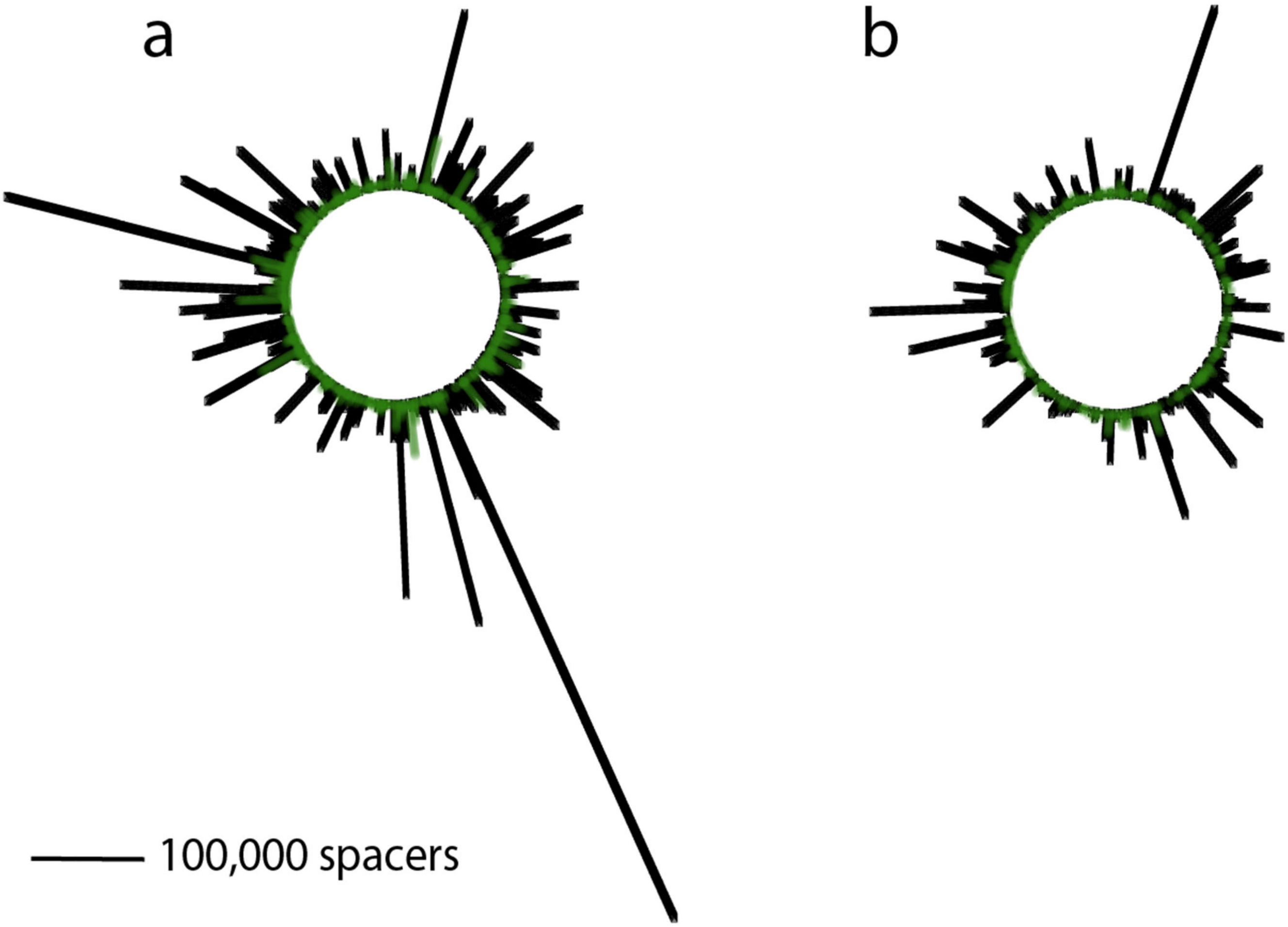
**Extended Data Figure 4. A model explaining the preference for spacer acquisition near *TerC* as compared to *TerA in E. coli* BL21-AI**

The DNA manipulation at the CRISPR region forms a replication fork stalling site, and leads to extensive spacer acquisition upstream to the CRISPR. While the clockwise fork is stalled at the CRISPR, the counterclockwise fork reaches the Ter region and is stalled at the respective Ter site, *TerC*, leading to extensive spacer acquisition upstream to *TerC*. Another factor that can contribute to the observed *TerC*/*TerA* bias may be that the clockwise replichore in *E. coli* (*oriC* to *TerA*) is longer than the counter clockwise one (*oriC* to *TerC*), leading the forks to stall at *TerC* more often than at *TerA*.



**Extended Data Figure 5. The protein product of T7 gene 5.9 inhibits spacer acquisition activity**

*E. coli* BL21-AI strains harboring pBAD-Cas1+2 and pBAD33-gp5.9 (lane 1) or pBAD33 vector control (lane 2) were grown overnight in the presence of inducers (0.4% L-arabinose). Gel shows PCR products amplified from the indicated cultures using primers annealing to the leader and to the fifth spacer of the CRISPR array. Results represent one of three independent experiments.



**Extended Data Figure 6. Distribution of protospacers across (A) pCtrl-*Chi* and (B) p*Chi* plasmids**

Circular representation of the 4.7kb plasmid is presented, with the inserted 4-*Chi* cluster present at the top-middle of the circle. Black bars indicate the number of PAM-derived spacers sequenced from each position; green bars represent non-PAM spacers. Scale bar indicates 100k spacers. Pooled protospacers from two replicates are presented for each panel.

## Extended Data Table 1a.

### Adaptation experiments with *E. coli* BL21-AI cells

Following overnight growth with or without induction of Cas1+2 cloned on pWUR plasmid, the CRISPR array was amplified and sequenced to determine the fraction of arrays that acquired a new spacer. Results with BL21-AI with an empty pWUR vector (without Cas1+2) are presented as a control.

| Sample | rep | # of reads spanning the CRISPR array | # of reads supporting unmodified parental array | # of reads supporting acquisition of a new spacer | % expanded arrays |
|---|---|---|---|---|---|
| **BL21-AI, no ara** | 1 | 25,718 | 25,163 | 555 | 2.16% |
| **BL21-AI, no ara** | 2 | 32,807 | 31,800 | 1,007 | 3.07% |
| **BL21-AI, 0.2% ara** | 1 | 28,188 | 17,438 | 10,750 | 38.14% |
| **BL21-AI, 0.2% ara** | 2 | 33,973 | 21,843 | 12,130 | 35.70% |
| **BL21-AI, empty vector, no ara** | 1 | 12,021 | 12,021 | 0 | 0% |
| **BL21-AI, empty vector, no ara** | 2 | 14,729 | 14,729 | 0 | 0% |
| **BL21-AI, empty vector, 0.2% ara** | 1 | 28,251 | 28,251 | 0 | 0% |
| **BL21-AI, empty vector, 0.2% ara** | 2 | 6,827 | 6,827 | 0 | 0% |

## Extended Data Table 1b.

### Identity of acquired spacers in *E. coli* BL21-AI cells

Following overnight growth with or without induction of Cas1+2, gel-separated expanded arrays were amplified and sequenced, to study the identity of newly acquired spacers in high resolution.

| Sample | rep | # new spacers sequenced | # spacers from chromosome | # spacers from plasmid | % spacers from plasmid | % spacers from genome |
|---|---|---|---|---|---|---|
| **BL21-AI, no ara** | 1 | 2,594,637 | 48,300 | 2,546,337 | 98.14% | 1.86% |
| **BL21-AI, no ara** | 2 | 2,056,397 | 35,911 | 2,020,486 | 98.25% | 1.75% |
| **BL21-AI, 0.2% ara** | 1 | 647,929 | 151,181 | 496,748 | 76.67% | 23.33% |
| **BL21-AI, 0.2% ara** | 2 | 851,824 | 190,791 | 661,033 | 77.60% | 22.40% |
| **BL21-AI pheA::TerB** | 1 | 2,937,147 | 46,015 | 2,891,132 | 98.43% | 1.57% |
| **BL21-AI pheA::TerB** | 2 | 3,400,210 | 44,748 | 3,355,462 | 98.68% | 1.32% |

## Extended Data Table 1c.

### Effect of antibiotics on adaptation levels

The Cas1+2 operon was induced in *E. coli* BL21-AI cells using 0.2% L-arabinose and 0.1mM IPTG overnight, in the presence of either nalidixic acid (50 µg/ml) or rifampicin (100 µg/ml). Following overnight induction, the CRISPR array was amplified and sequenced.

| Sample | rep | # of reads spanning the CRISPR array | # of reads supporting unmodified parental array | # of reads supporting acquisition of a new spacer | % expanded arrays |
|---|---|---|---|---|---|
| **BL21-AI + Nalidixic acid** | 1 | 71,941 | 71,800 | 141 | 0.20% |
| **BL21-AI + Nalidixic acid** | 2 | 77,774 | 77,714 | 60 | 0.08% |
| **BL21-AI + Rifampicin** | 1 | 36,976 | 34,145 | 2,831 | 7.66% |
| **BL21-AI + Rifampicin** | 2 | 38,702 | 28,147 | 10,555 | 27.27% |

## Extended Data Table 2a.

### Adaptation experiment with *dnaC2* temperature sensitive cells

*E. coli* K12 cells were transformed with a pBAD-Cas1+2 vector, in which the Cas1+2 operon is directly controlled by an arabinose-inducible promoter. Following overnight induction by 0.2% L-arabinose and 0.1mM IPTG, the CRISPR array was amplified and sequenced.

| Sample | rep | # of reads spanning the CRISPR array | # of reads supporting unmodified parental array | # of reads supporting acquisition of a new spacer | % expanded arrays |
|---|---|---|---|---|---|
| **K-12 ΔcasC 30°C** | 1 | 98,884 | 96,299 | 2,585 | 2.61% |
| **K-12 ΔcasC 30°C** | 2 | 117,522 | 115,030 | 2,492 | 2.12% |
| **K-12 ΔcasC, dnaC2 30°C** | 1 | 152,827 | 149,644 | 3,183 | 2.08% |
| **K-12 ΔcasC, dnaC2 30°C** | 2 | 100,125 | 98,053 | 2,072 | 2.07% |
| **K-12 ΔcasC, 39°C** | 1 | 87,036 | 83,688 | 3,348 | 3.85% |
| **K-12 ΔcasC, 39°C** | 2 | 86,580 | 82,474 | 4,106 | 4.74% |
| **K-12 ΔcasC, dnaC2 39°C** | 1 | 66,618 | 66,618 | 0 | 0.00% |
| **K-12 ΔcasC, dnaC2 39°C** | 2 | 60,325 | 60,321 | 4 | 0.01% |

## Extended Data Table 2b.

### Time course adaptation experiments with synchronously replicating *dnaC2* temperature sensitive cells

Temperature sensitive K-12ΔcasC*dnaC2* culture was transferred to 39°C for 70 minutes. Cas1+2 expression was then induced for 30 minutes using 0.2% L-arabinose and 0.1mM IPTG, and the culture was transferred to 30°C with continuous induction of Cas1+2. Culture was sampled at successive time points following synchronous replication initiation, and the CRISPR array was amplified and sequenced to determine the fraction of cells that acquired a new spacer. In addition, gel-separated expanded arrays were amplified and sequenced, to study the localization of spacers derived from the chromosome.

| | | Sequencing of the CRISPR array PCR product | | | | Direct sequencing of expanded arrays | | |
|---|---|---|---|---|---|---|---|---|
| **Sample** | **rep** | **# of reads spanning the CRISPR array** | **# of reads supporting unmodified parental array** | **# of reads supporting acquisition of a new spacer** | **% expanded arrays** | **# spacers from chromosome** | **# spacers from Ter region** | **% spacers from Ter** |
| **dnaC2 0 min, 30°** | 1 | 99,683 | 99,669 | 14 | 0.014% | 3,684 | 508 | 13.79% |
| **dnaC2 0 min, 30°** | 2 | 107,825 | 107,814 | 11 | 0.010% | 456 | 26 | 5.70% |
| **dnaC2 20 min, 30°** | 1 | 107,679 | 107,671 | 8 | 0.007% | 1,250 | 128 | 10.24% |
| **dnaC2 20 min, 30°** | 2 | 113,040 | 113,030 | 10 | 0.009% | 1,402 | 36 | 2.57% |
| **dnaC2 40 min, 30°** | 1 | 394,058 | 394,018 | 40 | 0.010% | 4,830 | 930 | 19.25% |
| **dnaC2 40 min, 30°** | 2 | 100,975 | 100,964 | 11 | 0.011% | 10,541 | 2,821 | 26.76% |
| **dnaC2 60 min, 30°** | 1 | 63,978 | 63,967 | 11 | 0.017% | 5,563 | 1,604 | 28.83% |
| **dnaC2 60 min, 30°** | 2 | 108,605 | 108,588 | 17 | 0.016% | 6,551 | 2,183 | 33.32% |
| **dnaC2 90 min, 30°** | 1 | 109,652 | 109,636 | 16 | 0.015% | 3,221 | 348 | 10.80% |
| **dnaC2 90 min, 30°** | 2 | 206,652 | 206,567 | 85 | 0.041% | 2,827 | 320 | 11.32% |
| **dnaC2 120 min, 30°** | 1 | 80,213 | 80,192 | 21 | 0.026% | 3,373 | 848 | 25.14% |
| **dnaC2 120 min, 30°** | 2 | 121,583 | 121,530 | 53 | 0.044% | 3,135 | 721 | 23.00% |

## Extended Data Table 3a.

### Adaptation experiments with *E. coli* BL21-AI Δ*recB*, Δ*recC*, Δ*recD* and ydhQ::I-*Sce*I cells

Following overnight growth without induction of Cas1+2, the CRISPR array was amplified and sequenced to determine the fraction of cells that acquired a new spacer. In addition, gel-separated expanded arrays were amplified and sequenced, to study the identity of newly acquired spacers in high resolution.

| | | Sequencing of the CRISPR array PCR product | | | | Direct sequencing of expanded arrays | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Sample** | **rep** | **# of reads spanning the CRISPR array** | **# of reads supporting unmodified parental array** | **# of reads supporting acquisition of a new spacer** | **% expanded arrays** | **# new spacers sequenced** | **# spacers from chromosome** | **# spacers from plasmid** | **% spacers from plasmid** | **% spacers from genome** |
| **BL21-AI ΔrecB, no ara** | 1 | 35,060 | 34,615 | 445 | 1.27% | 663,470 | 107,260 | 556,210 | 83.83% | 16.17% |
| **BL21-AI ΔrecB, no ara** | 2 | 36,116 | 35,778 | 338 | 0.94% | 441,290 | 75,260 | 366,030 | 82.95% | 17.05% |
| **BL21-AI ΔrecC, no ara** | 1 | 116,840 | 115,012 | 1,828 | 1.56% | 704,870 | 96,707 | 608,163 | 86.28% | 13.72% |
| **BL21-AI ΔrecC, no ara** | 2 | 132,549 | 130,724 | 1,825 | 1.38% | 507,844 | 55,057 | 452,787 | 89.16% | 10.84% |
| **BL21-AI ΔrecD, no ara** | 1 | 85,877 | 85,253 | 624 | 0.73% | 2,938,455 | 353,353 | 2,585,102 | 87.97% | 12.03% |
| **BL21-AI ΔrecD, no ara** | 2 | 90,498 | 89,802 | 696 | 0.77% | 4,437,733 | 1,405,158 | 3,032,575 | 68.34% | 31.66% |
| **BL21-AI ydhQ-l-*Sce*I site/pCasl2-IPTG/pBAD-I-*Sce*I, no ara** | 1 | 87,419 | 83,625 | 3,794 | 4.34% | 221,721 | 16,906 | 204,815 | 92.38% | 7.62% |
| **BL21-AI ydhQ-I-*Sce*I site/pCasl2-IPTG/pBAD-I-*Sce*I, no ara** | 2 | 89,357 | 86,745 | 2,612 | 2.92% | 192,597 | 15,995 | 176,642 | 91.72% | 8.28% |

## Extended Data Table 3b.

### Adaptation experiments with *E. coli* BL21-AI pCas1+2-*Chi*

Following overnight growth without induction of Cas1+2 from the p*Chi* plasmid (which contains a cluster of 4 counter clockwise *Chi* sites on a 50bp cassette inserted at position 1300 of the pCas plasmid) gel-separated expanded arrays were amplified and sequenced, to differentiate between spacers acquired from self and plasmid DNA. As a control a similar plasmid with a 50bp *Chi*-less insertion at the same position in the pCas plasmid was used.

**Direct sequencing of expanded arrays**

| Sample | rep | # new spacers sequenced | # spacers from chromosome | # spacers from plasmid | % spacers from plasmid | % spacers from genome |
|---|---|---|---|---|---|---|
| **BL21-AI pCtrl-*Chi*, no ara** | 1 | 4,221,820 | 42,055 | 4,179,765 | 99% | 1% |
| **BL21-AI pCtrl-*Chi*, no ara** | 2 | 5,743,373 | 50,345 | 5,693,028 | 99.12% | 0.88% |
| **BL21-AI p*Chi*, no ara** | 1 | 2,726,923 | 78,079 | 2,648,844 | 96.97% | 2.86% |
| **BL21-AI p*Chi*, no ara** | 2 | 2,841,509 | 87,106 | 2,754,403 | 96.74% | 3.06% |

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank Naama Barkai, Amos Tanay, Eran Mick, Shany Doron, Adi Stern, Tal Dagan and Maya Shamir for insightful discussion. We also thank the Skarstad group for providing the MG1655*dnaC2* strain, the Michel group for providing the JJC1819 strain, and Daniel Dar for assistance in Illumina sequencing. R.S. was supported, in part, by the ISF (personal grant 1303/12 and I-CORE grant 1796), the ERC-StG program (grant 260432), HFSP (grant RGP0011/2013), the Abisch-Frenkel foundation, the Pasteur-Weizmann council, the Minerva Foundation, and by a DIP grant from the Deutsche Forschungsgemeinschaft. U.Q. was supported, in part, by the European Research Council StG program (grant 336079), the Israel Science Foundation (grant 268/14), and the Israeli Ministry of Health (grant 9988-3). A.L. is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

## References

1. Terns MP, Terns RM. CRISPR-based adaptive immune systems. Current opinion in microbiology. 2011; 14:321–327. [PubMed: 21531607]
2. Westra ER, et al. The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet. 2012; 46:311–339. [PubMed: 23145983]
3. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature. 2012; 482:331–338. [PubMed: 22337052]
4. Koonin EV, Makarova KS. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA biology. 2013; 10:679–686. [PubMed: 23439366]
5. Sorek R, Lawrence CM, Wiedenheft B. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annu Rev Biochem. 2013; 82:237–266. [PubMed: 23495939]
6. Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Molecular cell. 2014; 54:234–244. [PubMed: 24766887]
7. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. 2012; 40:5569–5576. [PubMed: 22402487]
8. Nunez JK, et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nature structural & molecular biology. 2014
9. Swarts DC, Mosterd C, van Passel MW, Brouns SJ. CRISPR interference directs strand specific spacer acquisition. PLoS One. 2012; 7:e35888. [PubMed: 22558257]
10. Datsenko KA, et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun. 2012; 3:945. [PubMed: 22781758]
11. Diez-Villasenor C, Guzman NM, Almendros C, Garcia-Martinez J, Mojica FJ. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA biology. 2013; 10:792–802. [PubMed: 23445770]
12. Yosef I, et al. DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:14396–14401. [PubMed: 23940313]
13. Arslan Z, Hermanns V, Wurm R, Wagner R, Pul U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic acids research. 2014; 42:7884–7893. [PubMed: 24920831]
14. Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA biology. 2013; 10:716–725. [PubMed: 23619643]
15. Fineran PC, et al. Degenerate target sites mediate rapid primed CRISPR adaptation. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:E1629–1638. [PubMed: 24711427]

16. Skovgaard O, Bak M, Lobner-Olesen A, Tommerup N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome research. 2011; 21:1388–1393. [PubMed: 21555365]

17. Neylon C, Kralicek AV, Hill TM, Dixon NE. Replication termination in Escherichia coli: structure and antihelicase activity of the Tus-Ter complex. Microbiol Mol Biol Rev. 2005; 69:501–526. [PubMed: 16148308]

18. Waldminghaus T, Weigel C, Skarstad K. Replication fork movement and methylation govern SeqA binding to the Escherichia coli chromosome. Nucleic Acids Res. 2012; 40:5465–5476. [PubMed: 22373925]

19. Breier AM, Weier HU, Cozzarelli NR. Independence of replisomes in Escherichia coli chromosomal replication. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:3942–3947. [PubMed: 15738384]

20. del Solar G. Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev. 1998; 62:434–364. el al. [PubMed: 9618448]

21. Erdmann S, Le Moine Bauer S, Garrett RA. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. Molecular microbiology. 2014; 91:900–917. [PubMed: 24433295]

22. Smith GR. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev. 2012; 76:217–228. [PubMed: 22688812]

23. Dillingham MS, Kowalczykowski SC. RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev. 2008; 72:642–671. [PubMed: 19052323]

24. Kuzminov A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:8241–8246. [PubMed: 11459959]

25. Michel B, et al. Rescue of arrested replication forks by homologous recombination. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:8181–8188. [PubMed: 11459951]

26. Shee C, et al. Engineered proteins detect spontaneous DNA breakage in human and bacterial cells. eLife. 2013; 2:e01222. [PubMed: 24171103]

27. Babu M, et al. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol. 2011; 79:484–502. [PubMed: 21219465]

28. Lin, L. Study of bacteriophage T7 gene 5.9 and gene 5.5 PhD thesis. State University of New York; 1992.

29. Brouns SJ, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321:960–964. [PubMed: 18703739]

30. Baba T, et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2:1–11. [PubMed: 16738554]

31. Waldminghaus T, Weigel C, Skarstad K. Replication fork movement and methylation govern SeqA binding to the *Escherichia coli* chromosome. Nucleic Acids Res. 2012; 40:5465–5476. [PubMed: 22373925]

32. Bidnenko V, Ehrlich SD, Michel B. Replication fork collapse at replication terminator sequences. EMBO J. 2002; 21:3898–3907. [PubMed: 12110601]

33. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res. 2012; 40:5569–5576. [PubMed: 22402487]

34. Guzman LM, et al. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. J Bacteriol. 1995; 177:4121–4130. [PubMed: 7608087]

35. Tischer BK, et al. Two-step red-mediated recombination for versatile high-efficiency markerless DNA manipulation in *Escherichia coli*. Biotechniques. 2006; 40:191–197. [PubMed: 16526409]

36. Kitagawa M, et al. Complete set of ORF clones of *Escherichia coli* ASKA library (A Complete Set of E. coli K-12 ORF Archive): Unique Resources for Biological Research. DNA Res. 2005; 12:291–299. [PubMed: 16769691]

37. Sharan SK, et al. Recombineering: a homologous recombination-based method of genetic engineering. Nat Protoc. 2009; 4:206–223. [PubMed: 19180090]

38. Datta S, Costantino N, Court DL. A set of recombineering plasmids for gram-negative bacteria. Gene. 2006; 379:109–115. [PubMed: 16750601]

39. Svenningsen SL, et al. On the role of Cro in lambda prophage induction. Proc Natl Acad Sci U S A. 2005; 102:4465–4469. [PubMed: 15728734]

40. Yu D, et al. An efficient recombination system for chromosome engineering in *Escerichia coli*. Proc Natl Acad Sci U S A. 2000; 97:5978–83. [PubMed: 10811905]

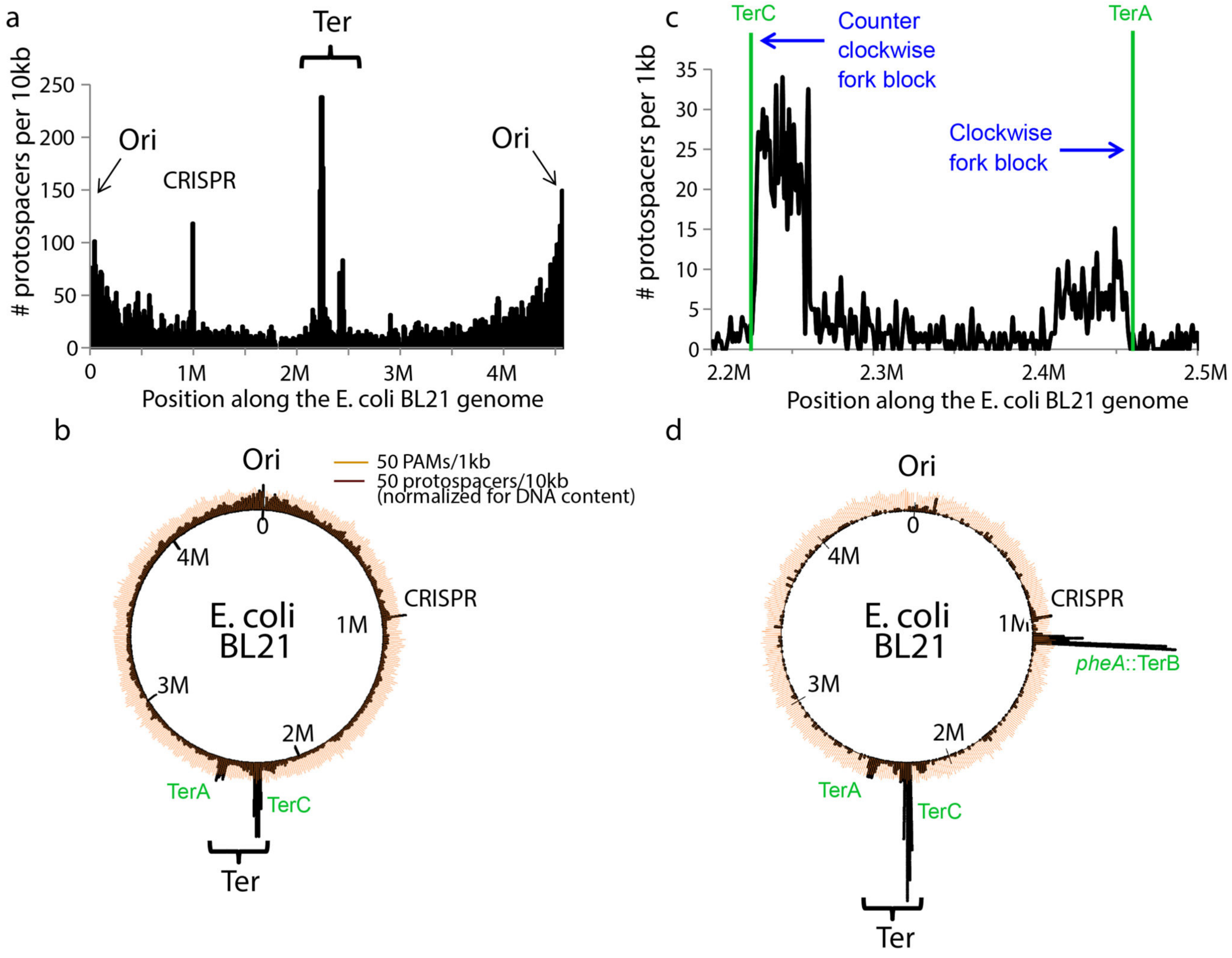


**Figure 1. Chromosome-scale hotspots for spacer acquisition**

(A) Distribution of protospacers across the *E. coli* BL21-AI genome. Protospacers were deduced from aligning new spacers, acquired into the CRISPR I array after 16 hours growth with no arabinose, to the bacterial genome. Only unique protospacers are presented, to avoid possible biases stemming from PCR amplification of the CRISPR array. Pooled protospacers from two replicates are presented. (B) Protospacer density across a circular representation of the *E. coli* genome, normalized to the DNA content of the culture. Dark brown, normalized protospacer numbers; orange, PAM density. (C) Protospacer distribution at the Ter region. Protospacer density is shown in 1kb windows. (D) Protospacer density in an *E. coli* BL21-AI in which the native 23bp-long *TerB* site was engineered into the *pheA* locus.

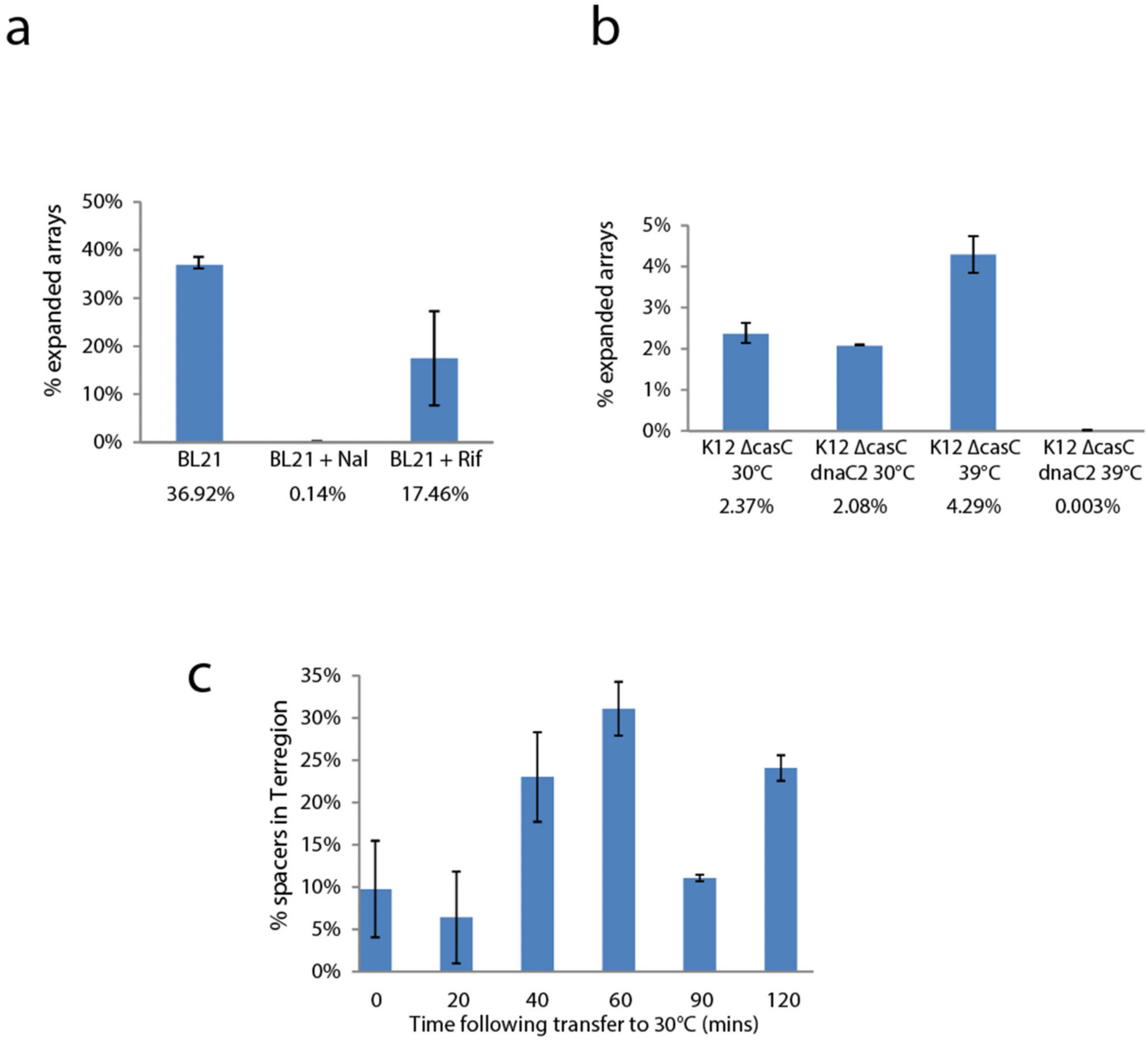


**Figure 2. Dependence of spacer acquisition in replication**

(A) Spacer acquisition rates in antibiotic-treated *E. coli* BL21-AI cells. Cells were grown at log phase 16 hours during Cas1+2 induction, with addition of the replication inhibitor nalidixic acid (Nal) or the transcription inhibitor rifampicin (Rif). (B) Spacer acquisition rates of K-12Δ*casCdnaC2* and an isogenic K-12Δ*casC* strains during overnight Cas1+2 induction. (C) Spacer acquisition patterns measured following transfer of K-12Δ*casCdnaC2* cells from 39°C to 30°C, during induction of Cas1+2. For all panels, average and error margins for two biological replicates are shown.

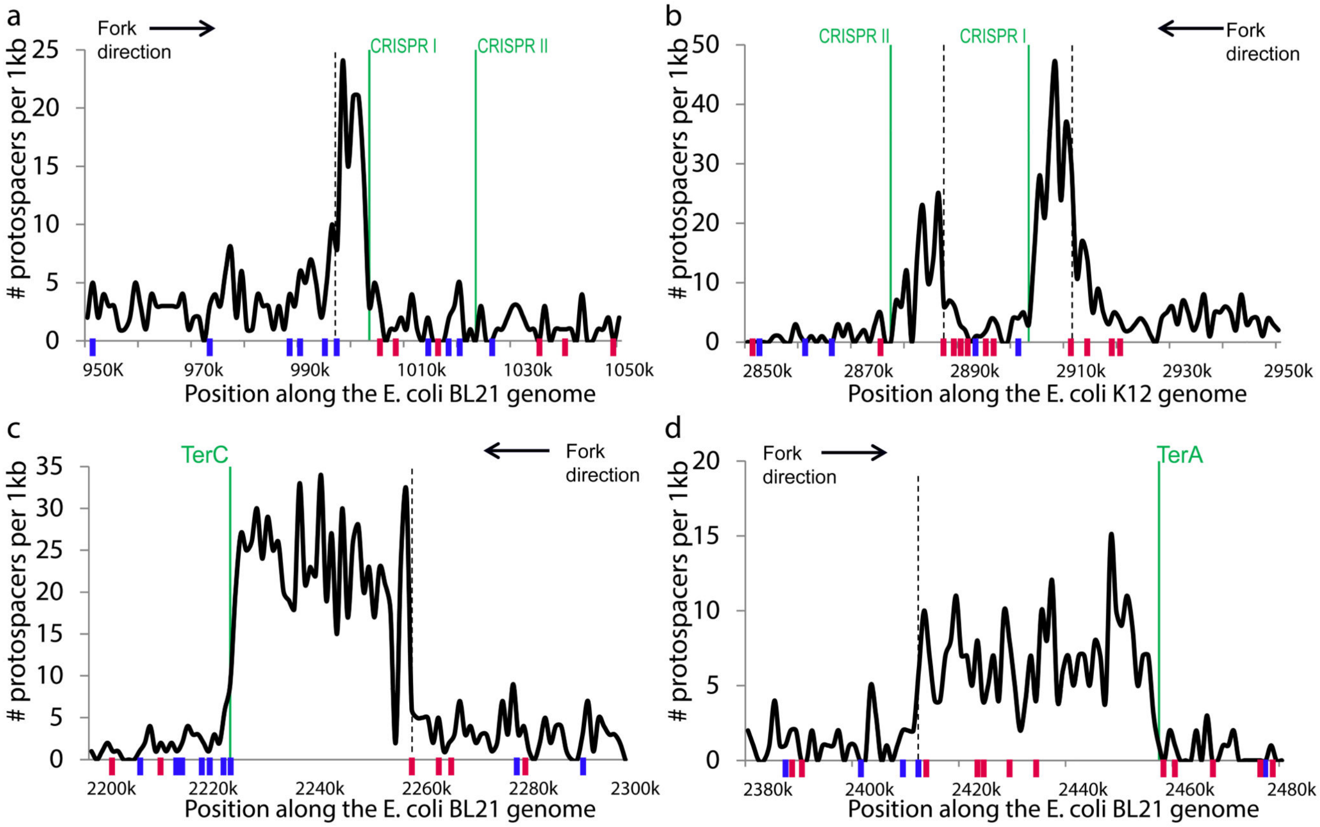


**Figure 3. *Chi* sites define boundaries of protospacer hotspots**

(A-D) Protospacer hotspot peaks. Each panel shows a 100kb window around a major hotspot for spacer acquisition. Short blue and red ticks mark positive- and negative-strand *Chi* sites, respectively. Green line mark a replication fork stalling site (*TerA*, *TerC*) or putative stalling site (CRISPR array). Dashed line marks the first properly oriented *Chi* site upstream relative to the fork stalling site. (A) The CRISPR region in *E. coli* BL21-AI. (B) The CRISPR region in *E. coli* K-12. (C) The *TerC* region and (D) the *TerA* region in *E. coli* BL21-AI. In panel C, the *Chi* site drawn at˜2260k represents a cluster of 3 consecutive *Chi* sites found in the same 1kb window.

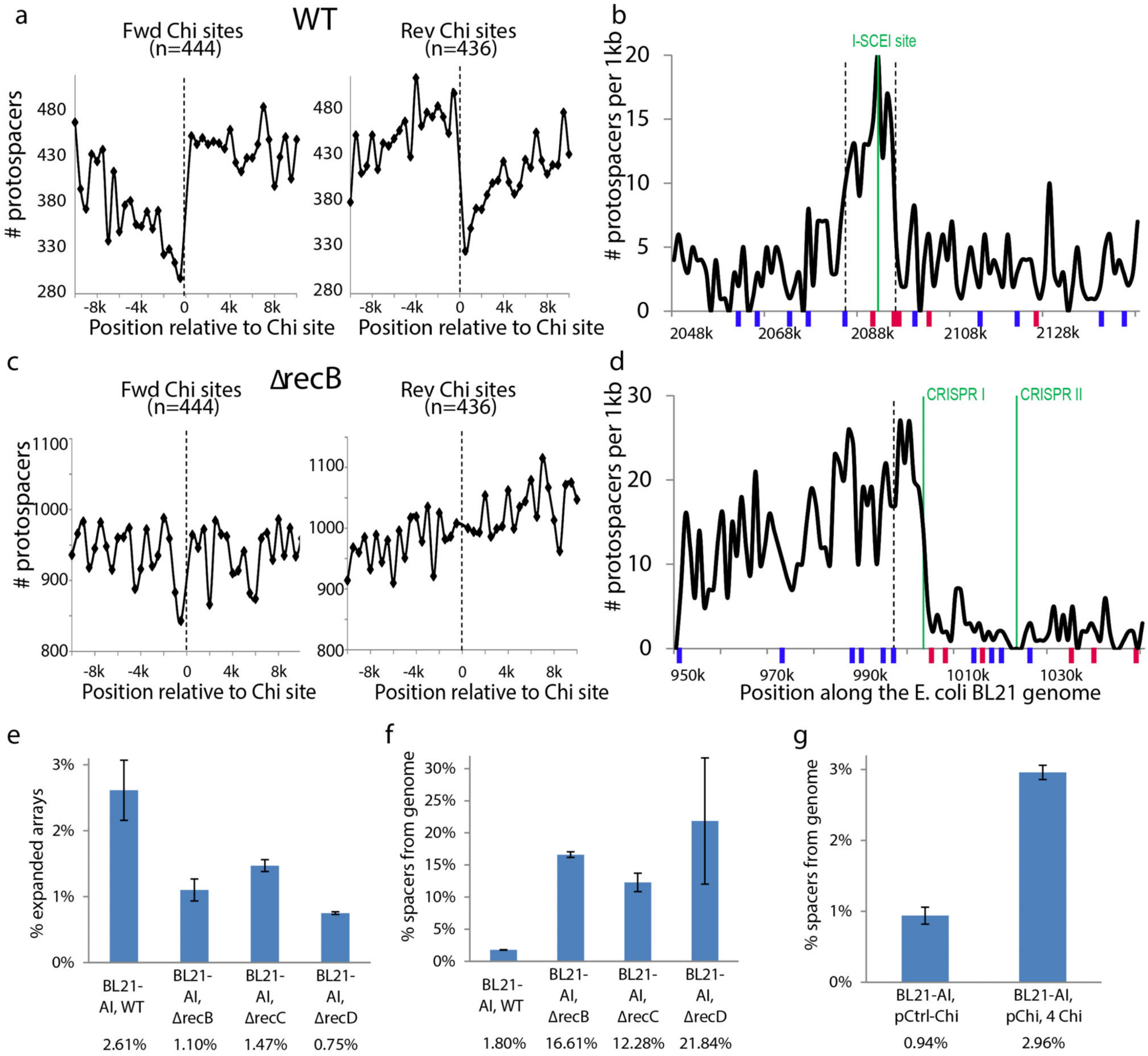


**Figure 4. Involvement of the dsDNA break repair machinery in defining spacer acquisition patterns**

(A) The overall number of protospacers around all *Chi* sites in *E. coli* BL21-AI, that are not included in the CRISPR region (950,000-1,050,000) or the Ter region (2M-2.5M), is shown in windows of 0.5 kb. (B) Protospacer hotspot peak resulting from a dsDNA break formed by the homing endonuclease I-*Sce*I.(C) The overall number of protospacers around all *Chi* sites that are not included in the CRISPR or the Ter regions in a BL21-AIΔ*recB* strain. (D) The protospacer hotspot at the CRISPR region in the BL21-AIΔ*recB* strain is not confined by a *Chi* site (compare to the same hotspot in the WT strain, Fig. 3A). (E) Adaption levels in WT BL21-AI and BL21-AIΔ*recB*, Δ*recC* or Δ*recD* strains following overnight growth without arabinose induction of Cas1+2. (F) Percent new spacers derived from the self chromosome in the experiment described in Panel E. (G) Percent new spacers derived from the self chromosome in the presence of a plasmid that contains a cluster of 4 *Chi* sites (p*Chi*) as compared to an identical plasmid that lacks *Chi* sites (pCtrl-*Chi*). For panels E-G, average and error margins for two biological replicates are shown.

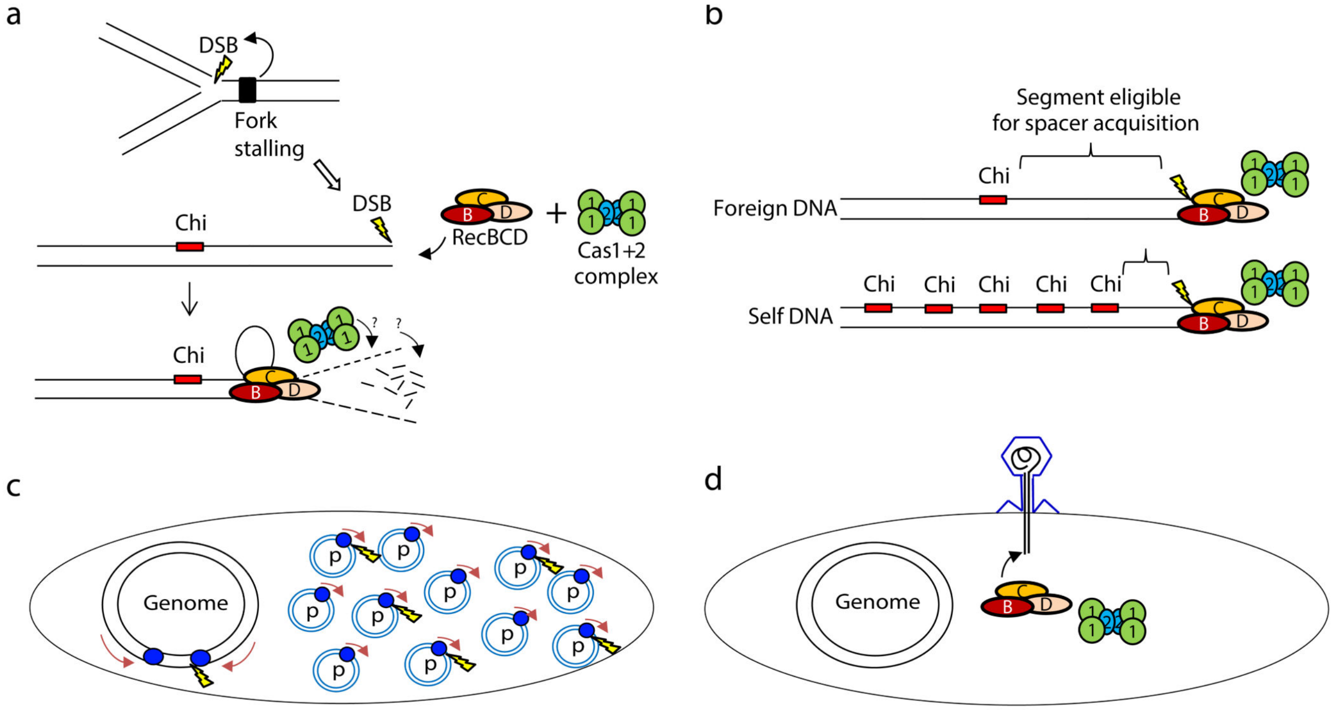


**Figure 5. A model explaining the preference for foreign DNA in spacer acquisition**

(A) RecBCD localizes to double strand DNA breaks (DSBs) and unwinds/degrades the DNA until reaching the nearest properly oriented *Chi* site. The RecBCD activity generates significant amounts of DNA “debris”, including short and long ssDNA fragments and degraded dsDNA, all of which may serve as substrates for spacer acquisition by Cas1+2. (B) High density of *Chi* sites on the chromosome reduces spacer acquisition from self DNA. On average, the 8bp-long *Chi* sites are found every 4.6kb on the *E. coli* chromosome, 14 times more often than on random DNA. When a DSB occurs on the chromosome, RecBCD DNA degradation activity will quickly be moderated by a nearby *Chi* site, but a similar DSB on a foreign DNA will lead to much more extensive DNA processing, providing more substrate for spacer acquisition. (C) Preference for spacer acquisition from high copy plasmids. In a replicating cell, most replication forks (blue circles) localize to the multiple copies of the plasmid. Since most DBSs occur during replication [23],[26] at stalled replication forks [24,25], plasmid DNA would become more amenable for spacer acquisition. (D) Most phages inject linear DNA into the infected cell. When such linear DNA is not protected, RecBCD will quickly degrade it, providing an intrinsic preference for spacer acquisition from phage DNA.