ARTICLE

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]§ & Rotem Sorek[1]§

CRISPR–Cas (clustered, regularly interspaced short palindromic repeats coupled with CRISPR-associated proteins) is a bacterial immunity system that protects against invading phages or plasmids. In the process of CRISPR adaptation, short pieces of DNA ('spacers') are acquired from foreign elements and integrated into the CRISPR array. So far, it has 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 double-stranded DNA break repair complex. Our results suggest that, in *Escherichia coli*, acquisition of new spacers largely depends on RecBCD-mediated processing of double-stranded DNA 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 both from high copy plasmids and from phages.

CRISPR–Cas is an adaptive defence system in bacteria and archaea that provides acquired immunity against phages and plasmids[1–6]. It comprises multiple *cas* genes, as well as 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 (or 'acquisition'), 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 CRISPR RNAs (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 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 plasmid DNA[7,8,11], suggesting an intrinsic preference for the Cas1–Cas2 complex to acquire spacers from the exogenous DNA. The mechanism by which the Cas1–Cas2 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–Cas2 in this system leads to spacer acquisition: that is, expansion of the chromosomally encoded CRISPR I array in *E. coli* BL21-AI[7]. Since this strain of *E. coli* harbours 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].

After 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, 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–Cas2 for 16 h, 36.92% (±1.2) of the sequenced arrays contained a new spacer. Conversely, in cultures where Cas1–Cas2 were not induced, 2.61% (±0.5) of the arrays contained a new spacer after 16 h of incubation, indicating that the leakage of Cas1–Cas2 transcription (as measured by RNA sequencing; Supplementary Table 1) still resulted in spacer acquisition in a

[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.

§These authors jointly supervised this work.

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 (4.7 kilobases (kb)) and its estimated copy number of 20–40, this represents 100- to 1,000-fold enrichment for acquisition of spacers from the plasmid, compared with what is expected by the DNA content in the cell. These results also show that lower expression of Cas1–Cas2 leads to higher specificity for exogenous DNA. Therefore, most of the analyses henceforth are based on spacers acquired in conditions in which Cas1–Cas2 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 and 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 of 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 *oriC* is replicated first, 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*-centred 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*-centred protospacer gradient (Fig. 1b).

The most striking protospacer hotspot was observed around the chromosomal replication terminus (*Ter*), in two major peaks showing approximately 7- to 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 of each fork stalling site and a relatively low, background protospacer density downstream of 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 the idea 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–Cas2. As a control, we applied the RNA polymerase inhibitor rifampicin, which blocks transcription in *E. coli* but allows DNA replication (this antibiotic does not interfere with transcription of Cas1–Cas2 by the T7 RNA polymerase). Application of nalidixic acid resulted in an almost complete

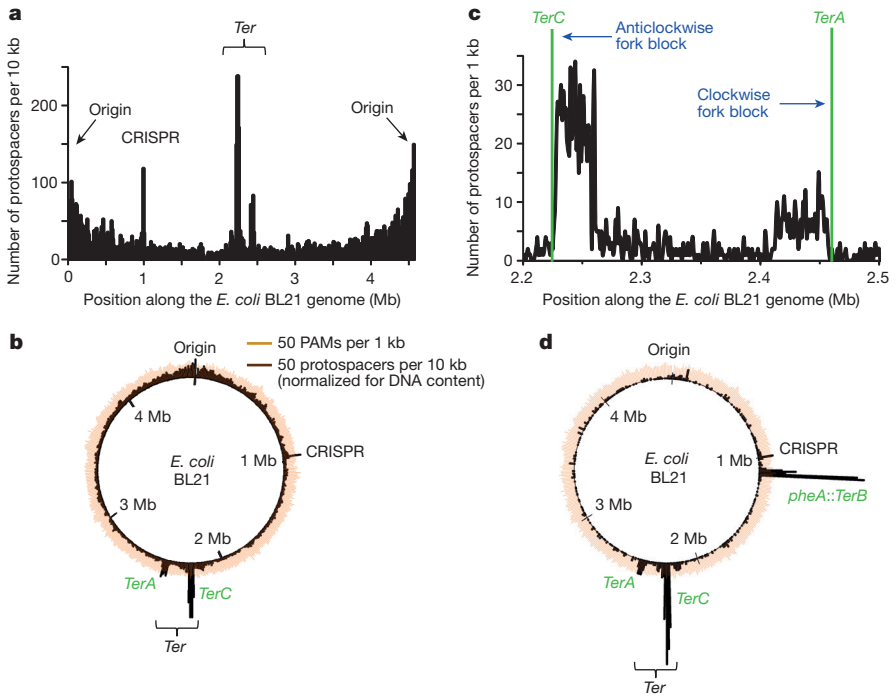


Figure 1 | Chromosome-scale hotspots for spacer acquisition. **a**, Distribution of protospacers across the *E. coli* BL21-AI genome. Protospacers were deduced from aligning new spacers, acquired into the CRISPR I array after 16 h growth with no arabinose, to the bacterial genome. Only unique protospacers 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 1-kb windows. **d**, Protospacer density in an *E. coli* BL21-AI in which the native 23 base pair (bp)-long *TerB* site was engineered into the *pheA* locus.

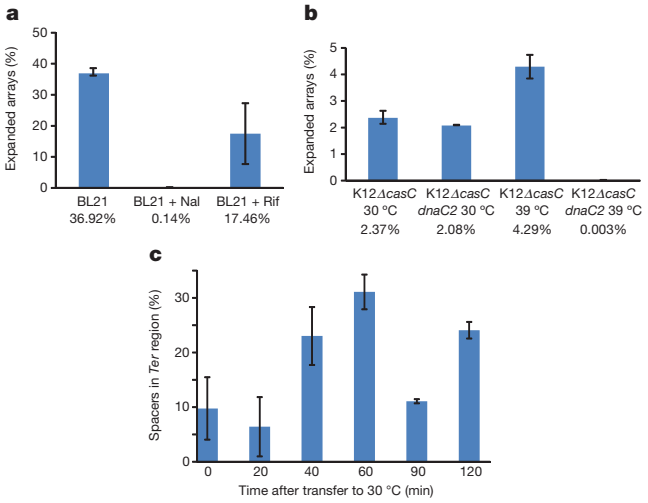


Figure 2 | Dependence of spacer acquisition on replication. **a**, Spacer acquisition rates in antibiotic-treated *E. coli* BL21-AI cells. Cells induced to express Cas1–Cas2 were grown for 16 h, with addition of the replication inhibitor nalidixic acid (Nal) or the transcription inhibitor rifampicin (Rif). **b**, Spacer acquisition rates of K-12*ΔcasCdnaC2* and an isogenic K-12*ΔcasC* strains during overnight Cas1–Cas2 induction. **c**, Spacer acquisition patterns measured after transfer of K-12*ΔcasCdnaC2* cells from 39 °C to 30 °C, during induction of Cas1–Cas2. For all panels, average and error margins for two biological replicates are shown.

elimination of spacer acquisition (164-fold reduction), but only an approximately twofold reduction in spacer acquisition rates was observed in the rifampicin-treated cells (Fig. 2a and Extended Data Table 1c), providing support to the hypothesis that spacer acquisition depends on DNA replication.

To substantiate these observations further, we examined the acquisition rates in *E. coli* K-12 cells carrying the temperature-sensitive allele *dnaC2* (ref. 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–Cas2 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 wild-type *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 and Extended Data Table 2a). These results further strengthen the hypothesis that Cas1–Cas2-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 min. Since in this temperature replication re-initiation is inhibited, after 70 min there are no more progressing forks in these cells. We then induced Cas1–Cas2 expression for 30 min, and transferred the culture to 30 °C, resulting in synchronized initiation of replication. At these conditions, it takes the replication forks on average about 60 min 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 min following synchronous replication initiation. Strikingly, the fraction of spacers derived from the *Ter* region gradually increased with the progression of the replication cycle, reaching 31% after 60 min (compared with only 6.4% at the 20 min time point; Fig. 2c, Extended Data Fig. 3 and Extended Data Table 2b). The pattern repeated itself in the second cycle of replication (90 and 120 min; 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–Cas2 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 travelling 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

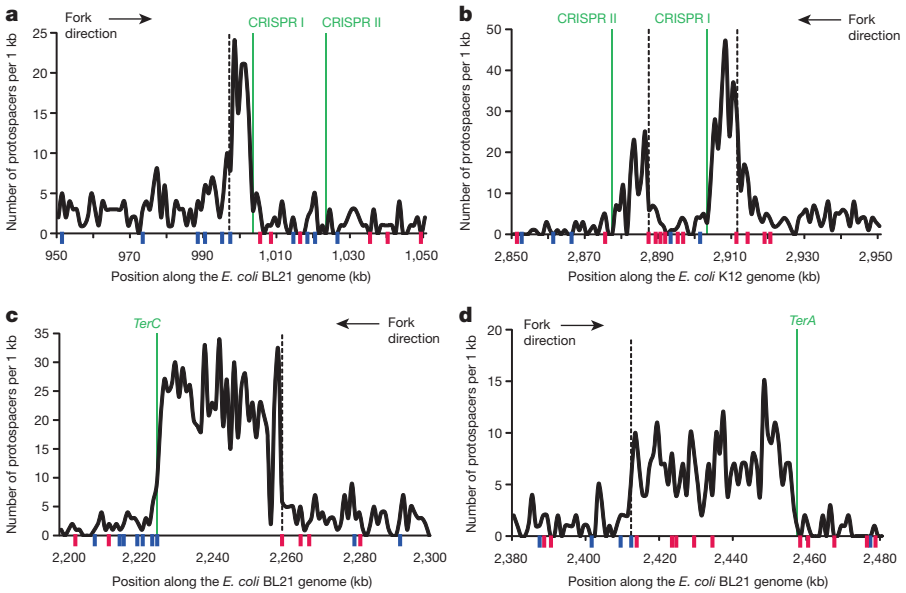


Figure 3 | Chi sites define boundaries of protospacer hotspots. **a**–**d**, Protospacer hotspot peaks. Each panel shows a 100 kb window around a major hotspot for spacer acquisition. Short blue and red ticks mark positive- and negative-strand Chi sites, respectively. Green lines mark a replication fork stalling site (*TerA*, *TerC*) or putative stalling site (CRISPR array). Dashed lines mark 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 **c**, the Chi site drawn at ~2,260 kb represents a cluster of three consecutive Chi sites found in the same 1 kb window.

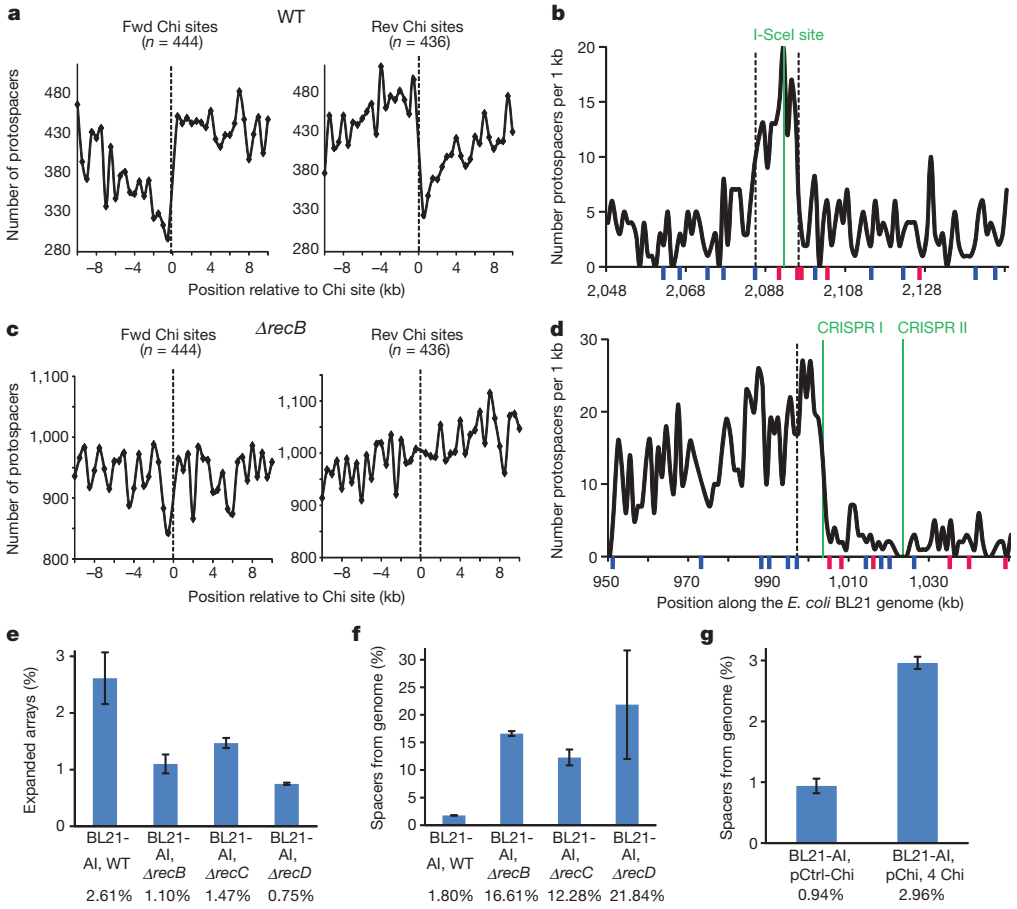


Figure 4 | Involvement of the dsDNA break repair machinery in defining spacer acquisition patterns. **a**, The overall number of protospacers around all Chi sites in *E. coli* BL21-AI, that are not included in the CRISPR region (950,000–1,050,000) or the *Ter* region (2–2.5 Mb), is shown in windows of 0.5 kb. WT, wild-type. **b**, Protospacer hotspot peak resulting from a dsDNA break formed by the homing endonuclease I-SceI. **c**, The overall number of protospacers around all Chi sites that are not included in the CRISPR or the *Ter* regions in a BL21-AI*ΔrecB* strain. **d**, The protospacer hotspot at the CRISPR region in the BL21-AI*ΔrecB* strain is not confined by a Chi site (compare with the same hotspot in the wild-type strain, Fig. 3a). **e**, Adaption levels in wild-type BL21-AI and BL21-AI*ΔrecB*, *ΔrecC* or *ΔrecD* strains after overnight growth without arabinose induction of Cas1–Cas2. **f**, Percentage of new spacers derived from the self chromosome in the experiment described in **e**. **g**, Percentage of new spacers derived from the self chromosome in the presence of a plasmid containing a cluster of four Chi sites (pChi) compared with an identical plasmid lacking Chi sites (pCtrl-Chi). For **e**–**g**, average and error margins for two biological replicates are shown.

travelling 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 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 of 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 that occurs 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 anti-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 and 50 kb (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 octamer 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 double-stranded DNA (dsDNA) break occurs, RecBCD localizes to the exposed end, and then unwinds

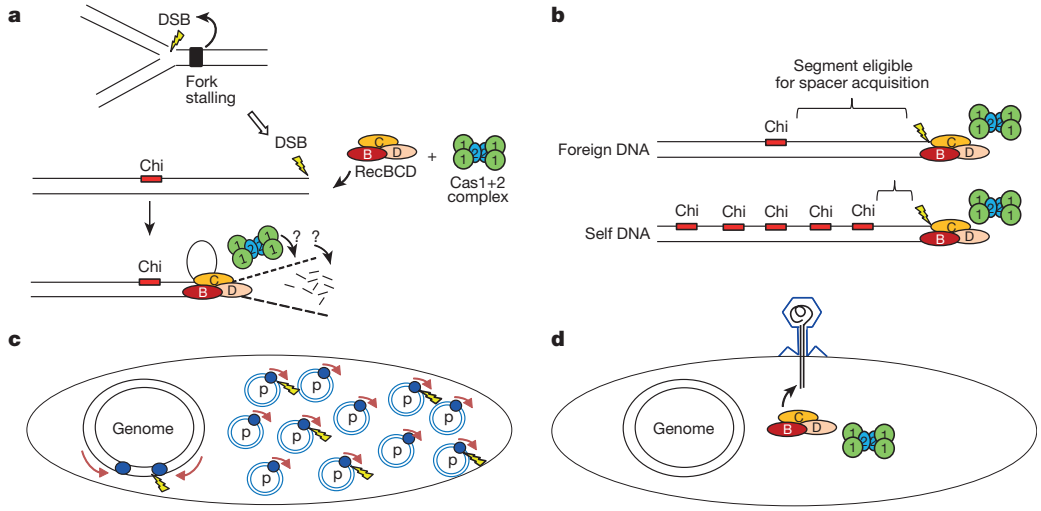


Figure 5 | A model explaining the preference for foreign DNA in spacer acquisition. **a**, RecBCD localizes to a dsDNA break (DSB) and unwinds/degrades the DNA until reaching the nearest properly oriented Chi site. The RecBCD activity generates significant amounts of DNA 'debris', including short and long ssDNA fragments and degraded dsDNA, all of which may serve as substrates for spacer acquisition by Cas1–Cas2. **b**, The high density of Chi sites on the chromosome reduces spacer acquisition from self DNA. On average, the 8-bp-long Chi sites are found every 4.6 kb on the *E. coli* chromosome, 14 times more often than on random DNA. When a dsDNA break occurs on the chromosome, RecBCD DNA degradation activity will quickly be moderated by a nearby Chi site, but a similar dsDNA break 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 dsDNA breaks 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.

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 single-stranded DNA 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 of 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 defence against phages[23]. Since Chi sites occur every ~5 kb in the *E. coli* genome, which is about 14 times more frequent than expected by chance, these sites were suggested as markers of bacterial self, preventing RecBCD from excessively degrading the chromosome after 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–Cas2 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 of 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 of Chi sites (Fig. 4a). The effect of this asymmetry was seen up to a distance of about 5–10 kb from the Chi site, consistent with an average distance of ~5 kb 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-SceI 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–Cas2 complex is directly fed from RecBCD DNA degradation products. Finally, Cas1 was shown to efficiently bind single-stranded DNA (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 and 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 compared with the wild-type strain (Fig. 4f and Extended Data Table 3a). These results show that CRISPR adaptation is partly dependent on the activity of the RecBCD dsDNA break repair complex, and that this activity is 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 compared with exogenous expression of a control protein (Extended Data Fig. 5).

It is noteworthy that in *recB* and *recC* deletions, the RecBCD complex is entirely non-functional, 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 a dsDNA break and a Chi site may be the source of 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–Cas2. In contrast, the high density of Chi sites on the bacterial chromosome serves for the relative avoidance of Cas1–Cas2 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 chromosome seen in the *recB*, *recC* and *recD* deletion strains conforms with the natural 14-fold enrichment of Chi sites on the chromosome. To examine further whether Chi sites limit spacer acquisition, we performed spacer acquisition experiments with a plasmid that was engineered to contain a cluster of four consecutive Chi sites. As expected, an increased preference for chromosomal DNA in spacer acquisition was measured for the Chi-containing plasmid (Fig. 4g, Extended Data Table 3b and 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–Cas2 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 owing 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 were successful, the extensive replication activity of the phage DNA would provide another anchor for spacer acquisition from phage.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 October 2014; accepted 9 February 2015.
Published online 13 April 2015.

1. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14,** 321–327 (2011).
2. Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46,** 311–339 (2012).
3. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482,** 331–338 (2012).
4. Koonin, E. V. & Makarova, K. S. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10,** 679–686 (2013).
5. Sorek, R., Lawrence, C. M. & Wiedenheft, B. CRISPR-mediated adaptive immune systems in Bacteria and Archaea. *Annu. Rev. Biochem.* **82,** 237–266 (2013).
6. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54,** 234–244 (2014).
7. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
8. Nunez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Struct. Mol. Biol.* **21,** 528–534 (2014).
9. Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7,** e35888 (2012).
10. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3,** 945 (2012).
11. Diez-Villasenor, C., Guzman, N. M., Almendros, C., Garcia-Martinez, J. & Mojica, F. J. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10,** 792–802 (2013).
12. Yosef, I. *et al.* DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc. Natl Acad. Sci. USA* **110,** 14396–14401 (2013).
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 Res.* **42,** 7884–7893 (2014).
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 Biol.* **10,** 716–725 (2013).
15. Fineran, P. C. *et al.* Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl Acad. Sci. USA* **111,** E1629–E1638 (2014).
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 Res.* **21,** 1388–1393 (2011).
17. Neylon, C., Kralicek, A. V., Hill, T. M. & Dixon, N. E. Replication termination in *Escherichia coli*: structure and antihelicase activity of the Tus-Ter complex. *Microbiol. Mol. Biol. Rev.* **69,** 501–526 (2005).
18. Waldminghaus, T., Weigel, C. & Skarstad, K. Replication fork movement and methylation govern SeqA binding to the *Escherichia coli* chromosome. *Nucleic Acids Res.* **40,** 5465–5476 (2012).
19. Breier, A. M., Weier, H. U. & Cozzarelli, N. R. Independence of replisomes in *Escherichia coli* chromosomal replication. *Proc. Natl Acad. Sci. USA* **102,** 3942–3947 (2005).
20. del Solar, G. *et al.* Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62,** 434–464 (1998).
21. Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of *Sulfolobus*. *Mol. Microbiol.* **91,** 900–917 (2014).
22. Smith, G. R. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiol. Mol. Biol. Rev.* **76,** 217–228 (2012).
23. Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72,** 642–671 (2008).
24. Kuzminov, A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc. Natl Acad. Sci. USA* **98,** 8241–8246 (2001).
25. Michel, B. *et al.* Rescue of arrested replication forks by homologous recombination. *Proc. Natl Acad. Sci. USA* **98,** 8181–8188 (2001).
26. Shee, C. *et al.* Engineered proteins detect spontaneous DNA breakage in human and bacterial cells. *eLife* **2,** e01222 (2013).
27. Babu, M. *et al.* A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.* **79,** 484–502 (2011).
28. Lin, L. *Study of Bacteriophage T7 Gene 5.9 and Gene 5.5.* PhD thesis, State Univ. New York (1992).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Barkai, A. Tanay, E. Mick, S. Doron, A. Stern, T. Dagan and M. Shamir for discussion. We also thank the Skarstad group for providing the MG1655*dnaC2* strain, the Michel group for providing the JJC1819 strain and D. Dar for assistance in Illumina sequencing. R.S. was supported, in part, by the Israel Science Foundation (personal grant 1303/12 and I-CORE grant 1796), the European Research Council Starting Grant programme (grant 260432), Human Frontier Science Program (grant RGP0011/2013), the Abisch-Frenkel foundation, the Pasteur-Weizmann Council, the Minerva Foundation, and by a Deutsch-Israelische Projektkooperation grant from the Deutsche Forschungsgemeinschaft. U.Q. was supported, in part, by the European Research Council Starting Grant programme (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.

Author 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. analysed data; A.L., M.G., U.Q. and R.S. wrote the manuscript.

Author Information RNA sequencing data are available in the National Center for Biotechnology Information Sequence Read Archive database under accession numbers SRX862155–SRX862158 in study SRP053013. Raw data of spacer sequences are accessible at http://www.weizmann.ac.il/molgen/Sorek/files/CRISPR_adaptation_2015/crispr_adaptation_2015_data.html. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to U.Q. (ehudq@post.tau.ac.il) or R.S. (rotem.sorek@weizmann.ac.il).

METHODS

No statistical methods were used to predetermine sample size.

Reagents, strains and plasmids. Luria–Bertani (LB) medium (10 g l^{-1} tryptone, 5 g l^{-1} yeast extract, 5 g l^{-1} NaCl) and agar were from Acumedia. Antibiotics and L-arabinose were from Calbiochem. Isopropyl-β-D-thiogalactopyranoside (IPTG) was from Bio-Lab. Calcium chloride ($CaCl_2$), sodium citrate (Na-citrate), restriction enzymes, T4 Polynucleotide Kinase 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. 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 done 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 (ref. 30) vector were both digested by SacI and SalI 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 four 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 NcoI-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 (ref. 30) vector were both digested by ScaI and SalI and ligated to yield pBAD33-gp5.9.

Strain construction using recombination-based genetic engineering (recombineering). All deletion mutants used for recombineering were obtained from the Keio collection[31]. BL21-AI *recB/C/D* deletion mutants were constructed as described previously[32]. Briefly, an overnight culture of BL21-AI/pSIM6 (ref. 33) was diluted 75-fold in 250 ml LB + 100 μg ml^{-1} ampicillin and aerated at 32 °C. When the attenuance, $D_{600\,nm}$, 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 4,600*g* at 0 °C for 10 min. The pellet was washed three times in ice-cold double-distilled H_2O, then resuspended in 200 μl ice-cold double distilled H_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^{-1} bacto-tryptone, 10 g l^{-1} yeast extract, 5 g l^{-1} NaCl) for 2 h in a 32 °C shaking water bath and inoculated on selection plates containing 25 μg ml^{-1} 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 the *dnaC2* allele linked to *Tn10* encoding tetracycline resistance marker[34], or *pheA* with the *TerB* site linked to spectinomycin[35]. P1 lysate was prepared as follows: overnight cultures of donor strain BL21-AI (for T7 RNA polymerase) or MG1655dnaC2 (for *dnaC2* allele)[34] or JJC1819 (for *pheA*::TerB-Spec)[35] were diluted 1:100 in 2.5 ml LB + 5 mM $CaCl_2$. After shaking for 1 h at 37 °C (or 30 °C for MG1655dnaC2), 0–100 μl phage P1 was added. Cultures were aerated for 1–3 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 0–100 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-SceI restriction site into the genome. A linear DNA containing the *Kan-sacB* cassette[36] for kanamycin resistance and sucrose sensitivity was amplified by PCR with oligonucleotides MG53F and MG53R that provided homology to a region downstream of the *ydhQ* gene. The *Kan-sacB* cassette was inserted into DY378 strain[37] by recombineering (as described above). Colonies that were found to be resistant to kanamycin and sensitive to sucrose (that is, 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 produced a short linear DNA containing the I-SceI restriction site with homology of 50 bp upstream and downstream of 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 the I-SceI restriction site immediately after the *ydhQ* stop codon. DNA from the resulting strain was sequence-verified for the presence of an intact I-SceI site.

CRISPR array size determination before acquisition assay. All strains underwent a preliminary validation step aimed at eliminating acquisition before induction: *E. coli* BL21-AI or K-12 harbouring pCas1+2 or pBAD-Cas1+2 plasmids, respectively, were spread on LB + 50 μg ml^{-1} streptomycin or 100 μg ml^{-1} ampicillin + 0.2% (w/v) glucose plates and incubated overnight at 37 °C or 30 °C (for K12*ΔcasCdnaC2*). 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-AI*ΔpheA::terB*, or BL21-AI*ΔrecB/C/D* strains harbouring pCas1+2 plasmid, or BL21-AI strain harbouring pWURV2 plasmid, or K-12*ΔcasC* T7RNAP strain harbouring pBAD-Cas1+2 plasmid, or BL21-AI *ydhQ*-I-SceI site strain harbouring pCas1+2-IPTG and pBAD-I-SceI[38] plasmids, or BL21-AI strain harbouring pChi or pCtrl-Chi plasmids, and BL21-AI strain harbouring pCas1+2 and pBAD33-gp5.9 plasmids were inoculated in LB medium containing 50 μg ml^{-1} streptomycin + 0.2% (w/v) glucose for BL21-AI strains carrying a single plasmid, or 100 μg ml^{-1} ampicillin + 0.2% (w/v) glucose for K12 strain, or 100 μg ml^{-1} ampicillin + 50 μg ml^{-1} streptomycin + 0.2% (w/v) glucose for BL21-AI ydhQ-I-SceI site/pCas1+2-IPTG/pBAD-I-SceI strain, or 200 μg ml^{-1} ampicillin + 35 μg ml^{-1} 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% (w/v) L-arabinose + 0.1 mM IPTG for pCas1+2, pChi and pCtrl-Chi harbouring strains, or 0.2% (w/v) L-arabinose for pBAD-Cas1+2 harbouring strains, or 0.02 mM IPTG and 0% L-arabinose for pCas1+2-IPTG and pBAD-I-SceI harbouring strains, or 0.4% (w/v) L-arabinose for pCas1+2 and pBAD33-gp5.9 harbouring strains. Cultures were aerated at 37 °C for an 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. A single colony of BL21-AI/pCas1+2 was inoculated in LB medium containing 50 μg ml^{-1} streptomycin + 0.2% (w/v) glucose and aerated at 37 °C for 16 h. The overnight cultures were diluted 1:600 in LB medium containing 50 μg ml^{-1} streptomycin with or without 0.2% (w/v) L-arabinose + 0.1 mM IPTG and aerated at 37 °C. Once cultures reached a $D_{600\,nm}$ of 0.25, cells were centrifuged in a microcentrifuge for 10 min at 13,000*g* and resuspended in LB medium containing 50 μg ml^{-1} streptomycin or 50 μg ml^{-1} nalidixic acid or 100 μg ml^{-1} rifampicin with or without 0.2% (w/v) L-arabinose + 0.1 mM IPTG. Cultures were aerated for 16 h at 37 °C, lysed and served as template for PCRs using primers OA1F/IY130R (PCR1) and RE10RD/IY230R (PCR2) for amplifying BL21-AI CRISPR array I.

Acquisition assay in replication-deficient strains. A single colony of K-12*ΔcasC* (control) or K-12*ΔcasCdnaC2* harbouring pBAD-Cas1+2 was inoculated in LB medium containing 100 μg ml^{-1} ampicillin + 0.2% (w/v) glucose and aerated at 30 °C, for 16 h. The overnight cultures were diluted 1:600 in LB medium containing 100 μg ml^{-1} ampicillin + 0.2% (w/v) 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. A single colony of K-12*ΔcasC* (control) or K-12*ΔcasCdnaC2* harbouring pBAD-Cas1+2 was inoculated in LB medium containing 100 μg ml^{-1} ampicillin + 0.2% (w/v) glucose and aerated at 30 °C, for 16 h. The overnight cultures were diluted 1:600 in LB medium containing 100 μg ml^{-1} ampicillin + 0.2% (w/v) glucose and aerated at 30 °C until $D_{600\,nm}$ reached 0.25. Cultures were then split into six tubes and transferred to non-permissive temperature (39 °C). After 70 min, induction of Cas1–Cas2

was performed: cells were centrifuged in a standard centrifuge (4,600*g*, 10 min), resuspended in LB medium containing 100 μg ml^{-1} ampicillin + 0.2% (w/v) L-arabinose and aerated for an additional 30 min at 39 °C. Replication was then initiated by aerating the split cultures at 30 °C for 0, 20, 40, 60, 90 and 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. One millilitre of each culture was centrifuged in a microcentrifuge for 1 min at 13,000*g* and resuspended in 100 μl LB medium. The concentrated culture underwent fast freeze in liquid nitrogen, was 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*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 sequencing. A single colony of *E. coli* BL21-AI strain harbouring pCas1+2 plasmid was inoculated in LB medium containing 50 μg ml^{-1} streptomycin + 0.2% (w/v) 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% (w/v) L-arabinose + 0.1 mM IPTG. After overnight growth, 15 ml from each culture was centrifuged in a standard centrifuge (4,600*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^{-1} lysozyme (Sigma-Aldrich catalogue number L6876-1G) in Tris 10 mM EDTA 1 mM pH 8.0, and total nucleotides were extracted using the Tri-Reagent protocol, according to the manufacturer's instructions (Molecular Research Center, catalogue number TR118). TURBO DNA-free Kit was used to eliminate DNA from the sample, according to the manufacturer's instructions (Life Technologies – Ambion catalogue number AM1907). Enrichment for messenger RNA (mRNA) was accomplished by using the Ribo-Zero rRNA Removal Kits (Illumina-Epicentre catalogue number MRZB12424). The enriched mRNA sample was then further purified using Agencourt AMPure XP magnetic beads (Beckman Coulter catalogue number A63881). Purified bacterial mRNA was then used as the starting material for the preparation of cDNA libraries for next-generation sequencing using a NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB catalogue number E7420S). The NEBNext multiplex oligonucleotides for Illumina Index primer set1 (NEB catalogue number E7335S) were used as the adapters for the library.

Total DNA purification. Overnight cultures of *E. coli* BL21-AI or K-12*ΔcasC* T7RNAP harbouring pCas1+2 or pBAD-Cas1+2 plasmid, respectively, were diluted 1:600 and aerated for 16 h at 37 °C in LB medium containing 50 μg ml^{-1} streptomycin or 100 μg ml^{-1} ampicillin + 0.2% (w/v) glucose. These overnight cultures were then diluted 1:600 in LB medium containing 50 μg ml^{-1} streptomycin or 100 μg ml^{-1} ampicillin with 0.2% (w/v) L-arabinose + 0.1 mM IPTG or without inducers and aerated at 37 °C. Once cultures reached $D_{600\,nm}$ = 0.5–0.6, 3 ml were removed and used for total DNA purification using a 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 performed for 5 min. Half of the PCR1 content (20 μl) was purified using the DNA clean-up kit and were used for standard library preparation procedures followed by deep sequencing (MiSeq), while the other half (20 μl) was loaded on a 2% (w/v) 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. One nanogram from the extracted band served as a template for the PCR2 reaction aimed at amplifying 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% (w/v) 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 the manufacturer's instructions. Several samples were multiplexed together in the same sequencing run. Demultiplexing was performed on the basis of different Illumina barcodes and on the basis of the 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; for strain BL21-AI (for which the genomic sequence is unavailable), the *E. coli* BL21-Gold(DE3)pLysS AG was used (Refseq accession NC_012947.1).

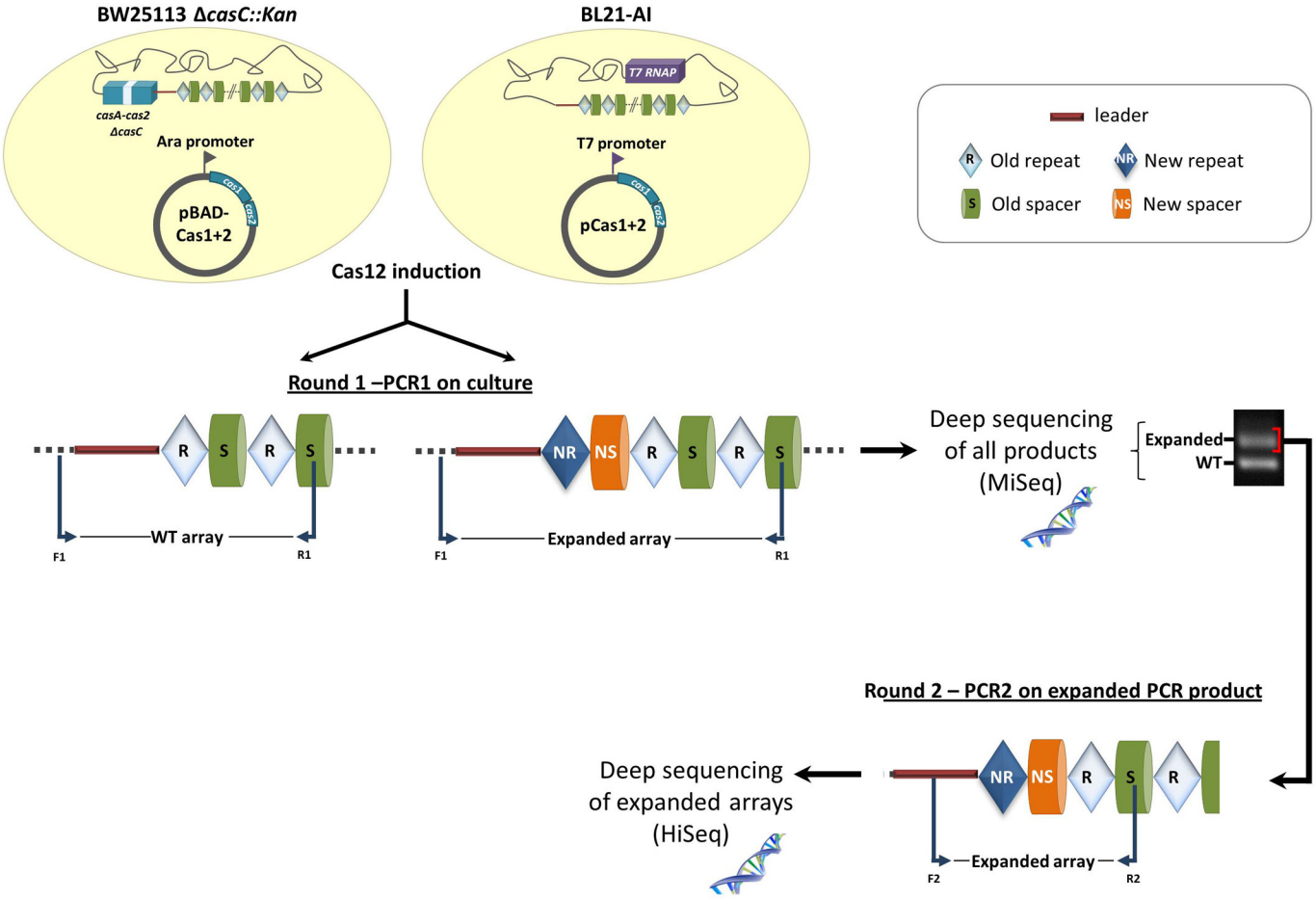
New spacer insertions were called on the basis of sequence alignments of the resulting reads. For round 1 of the PCR (Extended Data Fig. 1), alignments supporting non-acquisition events were also recorded to quantify acquisition level. If the sequence read was fully mapped to the parental CRISPR locus in the leader-proximal side, a non-acquisition event was inferred. New acquisition events were inferred if the read alignment began 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 (for example, spacers mapped to repetitive rRNA genes were discarded). If a spacer was mapped equally well both to the genome and the pCAS plasmid, only the plasmid protospacer position was used.

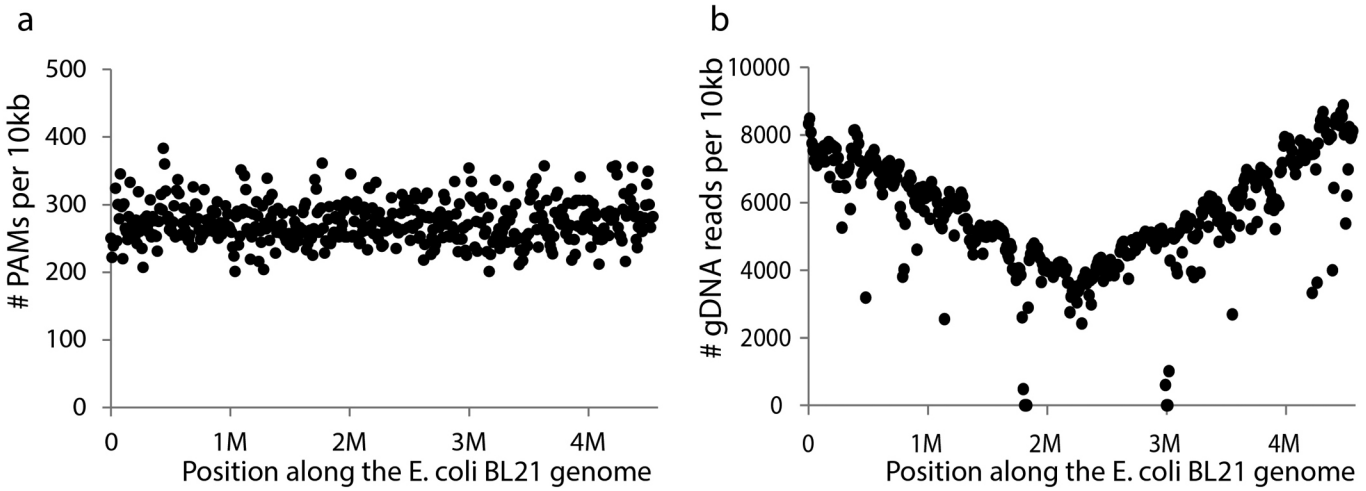
For the plots of protospacer distribution and hotspots (except for the plot in Extended Data Fig. 3), protospacer positions were recorded only once (meaning that if there were multiple spacers hitting the exact same position, the position was considered only once). This procedure was done to avoid biases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

Perl and R scripts were used for data analysis. Data visualization and statistical analysis used Microsoft Excel and R, including the R circular package (http://cran.r-project.org/web/packages/circular/circular.pdf) for Fig. 1 and Extended Data Fig. 4.

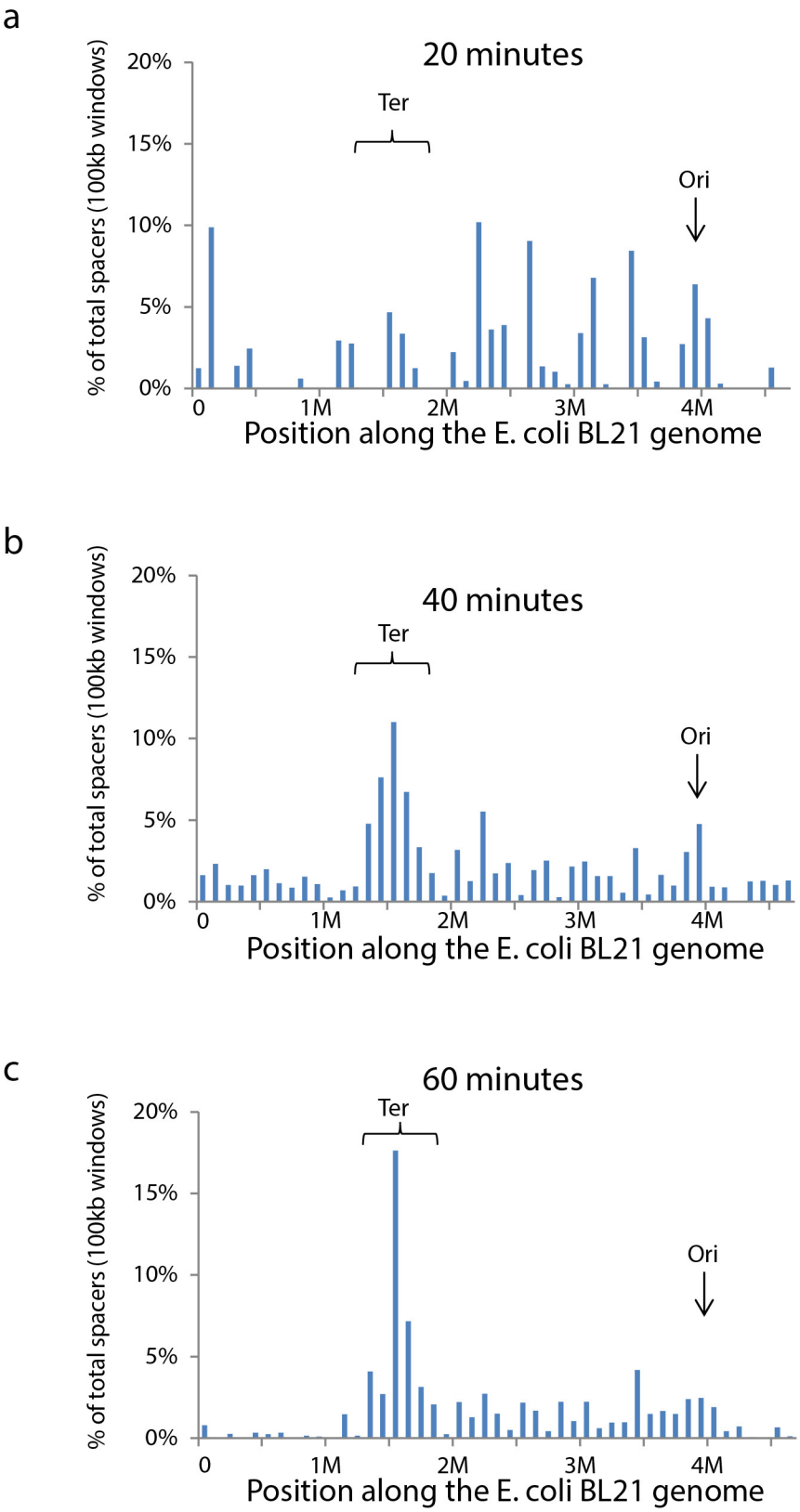
29. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–964 (2008).
30. Guzman, L. M. *et al.* Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177,** 4121–4130 (1995).
31. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 1–11 (2006).
32. Sharan, S. K. *et al.* Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols* **4,** 206–223 (2009).
33. Datta, S., Costantino, N. & Court, D. L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379,** 109–115 (2006).
34. Waldminghaus, T., Weigel, C. & Skarstad, K. Replication fork movement and methylation govern SeqA binding to the *Escherichia coli* chromosome. *Nucleic Acids Res.* **40,** 5465–5476 (2012).
35. Bidnenko, V., Ehrlich, S. D. & Michel, B. Replication fork collapse at replication terminator sequences. *EMBO J.* **21,** 3898–3907 (2002).
36. Svenningsen, S. L. *et al.* On the role of Cro in lambda prophage induction. *Proc. Natl Acad. Sci. USA* **102,** 4465–4469 (2005).
37. Yu, D. *et al.* An efficient recombination system for chromosome engineering in *Escerichia coli. Proc. Natl Acad. Sci. USA* **97,** 5978–5983 (2000).
38. Tischer, B. K. *et al.* Two-step red-mediated recombination for versatile high-efficiency markerless DNA manipulation in *Escherichia coli. Biotechniques* **40,** 191–197 (2006).



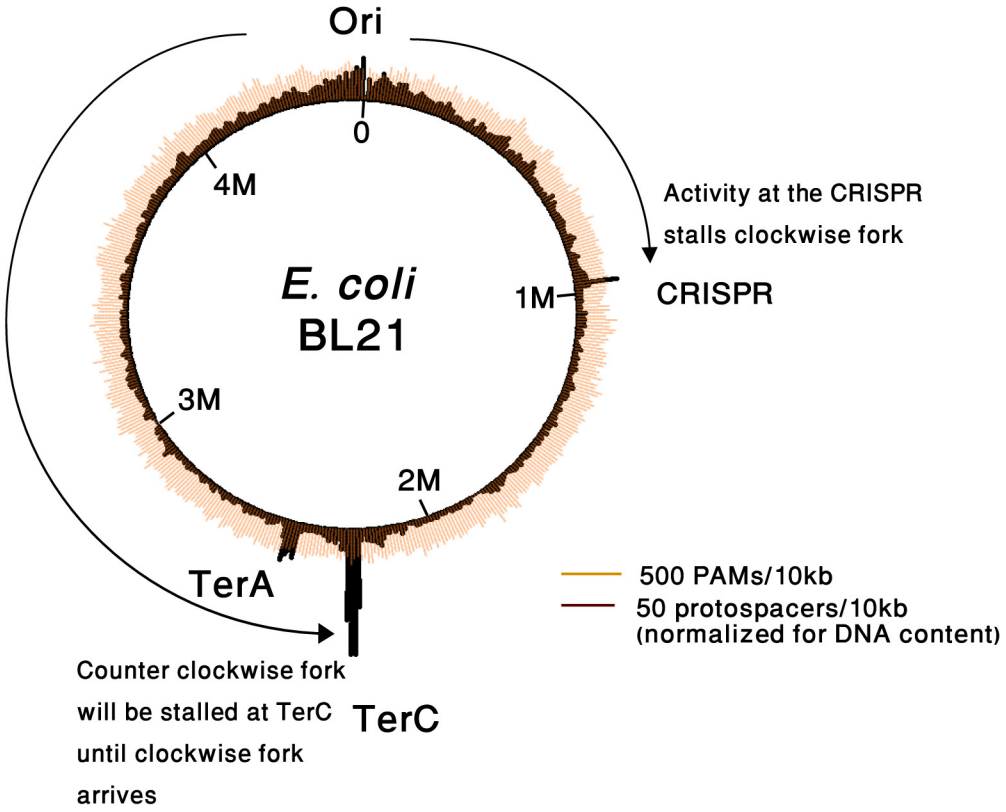
Extended Data Figure 1 | Graphic overview of the procedure for characterizing the frequency and sequence of newly acquired spacers. DNA from cultures of either *E. coli* K-12 (left) or *E. coli* BL21-AI (right) strains expressing Cas1–Cas2 from two different plasmids were used as templates for PCR. Round 1 was used to determine the frequency of spacer acquisition by comparing occurrences of expanded arrays to wild-type (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.



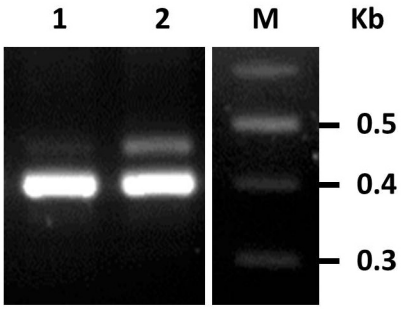
Extended Data Figure 2 | PAMs and DNA content along the *E. coli* BL21-AI genome. **a**, Distribution of PAM (AAG) sequences. Each data point represents the number of PAMs in a window of 10 kb. **b**, DNA content of a culture growing in log phase. Genomic DNA was extracted from *E. coli* BL21-AI cells 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 number NC_012947). Areas where few 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).



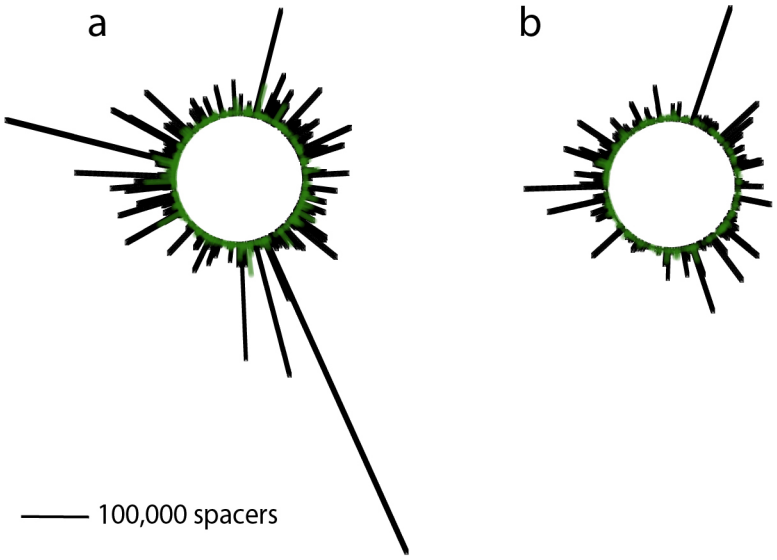
Extended Data Figure 3 | Distribution of newly acquired spacers on the genome during synchronized replication. *E. coli* K-12*ΔcasCdnaC2* cells were transferred from 39 °C (replication restrictive temperature) to 30 °C (replication permissive). Cas1–Cas2 were induced in these cells 30 min before the transfer to 30 °C and during the growth in 30 °C. Newly acquired spacers were sequenced at the given time points: **a**, following 20 min; **b**, following 40 min; **c**, following 60 min from replication initiation. The positions of the newly acquired spacers in windows of 100 kb are shown, 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* compared with *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 of the CRISPR. While the clockwise fork is stalled at the CRISPR, the anticlockwise fork reaches the *Ter* region and is stalled at the respective *Ter* site, *TerC*, leading to extensive spacer acquisition upstream of *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 anticlockwise 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 harbouring 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 the plasmids. **a**, Distribution across pCtrl-Chi; **b**, distribution across pChi plasmids. Circular representation of the 4.7 kb plasmid is presented, with the inserted 4-Chi cluster present at the top of the circle. Black bars indicate the number of PAM-derived spacers sequenced from each position; green bars represent non-PAM spacers. Scale bar, 100,000 spacers. Pooled protospacers from two replicates are presented for each panel.

Extended Data Table 1 | Spacer acquisition in normal and perturbed conditions

a.

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%

b.

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%

c.

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%

a, Adaptation experiments with *E. coli* BL21-AI cells. After overnight growth with or without induction of Cas1–Cas2 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–Cas2) are presented as a control. **b**, Identity of acquired spacers in *E. coli* BL21-AI cells. After overnight growth with or without induction of Cas1–Cas2, gel-separated expanded arrays were amplified and sequenced, to study the identity of newly acquired spacers in high resolution. **c**, Effect of antibiotics on adaptation levels. The Cas1–Cas2 operon was induced in *E. coli* BL21-AI cells using 0.2% L-arabinose and 0.1 mM IPTG overnight, in the presence of either nalidixic acid (50 $\mu g\,ml^{-1}$) or rifampicin (100 $\mu g\,ml^{-1}$). After overnight induction, the CRISPR array was amplified and sequenced.

Extended Data Table 2 | Replication-dependent spacer acquisition

a.

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%

b.

		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 chromo-some	# 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%

a, Adaptation experiment with *dnaC2* temperature-sensitive cells. *E. coli* K12 cells were transformed with a pBAD-Cas1–Cas2 vector, in which the Cas1–Cas2 operon was directly controlled by an arabinose-inducible promoter. After overnight induction by 0.2% L-arabinose and 0.1 mM IPTG, the CRISPR array was amplified and sequenced. **b,** Time-course adaptation experiments with synchronously replicating *dnaC2* temperature-sensitive cells. Temperature-sensitive K-12 *ΔcasC dnaC2* culture was transferred to 39 °C for 70 min. Cas1–Cas2 expression was then induced for 30 min using 0.2% L-arabinose and 0.1 mM IPTG, and the culture was transferred to 30 °C with continuous induction of Cas1–Cas2. The culture was sampled at successive time points after 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.

Extended Data Table 3 | Involvement of the DNA repair machinery in spacer acquisition

a.

		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**	**% expand-ed arrays**	**# new spacers sequen-ced**	**# spacers from chromo-some**	**# 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-I-*Sce*I site/pCas12-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/pCas12-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%

b.

		Direct sequencing of expanded arrays				
Sample	**rep**	**# new spacers sequen-ced**	**# spacers from chromo-some**	**# 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%

a, Adaptation experiments with *E. coli* BL21-AI*ΔrecB, ΔrecC, ΔrecD* and ydhQ::I-SceI cells. After overnight growth without induction of Cas1–Cas2, 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. **b**, Adaptation experiments with *E. coli* BL21-AI pCas1+2-Chi. After overnight growth without induction of Cas1–Cas2 from the pChi plasmid (which contained a cluster of four anticlockwise Chi sites on a 50 bp cassette inserted at position 1,300 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 50-bp Chi-less insertion at the same position in the pCas plasmid was used.