DNA targeting and interference by a bacterial Argonaute nuclease

DNA targeting and interference by a bacterial Argonaute nuclease

Anton Kuzmenko[1,2*], Anastasiya Oguienko[1,4], Daria Esyunina[1,4], Denis Yudin[1,4,5], Mayya Petrova[1], Alina Kudinova[1], Olga Maslova[1], Maria Ninova[2], Sergei Ryazansky[1], David Leach[3], Alexei A. Aravin[1,2*], Andrey Kulbachinskiy[1*]

[1] Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia

[2] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA

[3] Institute of Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3FF, UK

[4]These authors contributed equally: Anastasiya Oguienko, Daria Esyunina, Denis Yudin

[5] Present address: Department of Biology, Institute of Molecular Biology and Biophysics, ETH Zurich, Otto-Stern-Weg 5, CH-8093 Zurich, Switzerland

* To whom correspondence should be addressed:
AKuz, dr.ant.kuz@gmail.com; AAA, aaa@caltech.edu; AKul, avkulb@yandex.ru

Members of the conserved Argonaute protein family use small RNA guides to find their mRNA targets to regulate gene expression and suppress mobile genetic elements in eukaryotes [1,2]. Argonautes are also present in many bacterial and archaeal species [3-5]. Unlike eukaryotic proteins, several studied prokaryotic Argonautes use small DNA guides to cleave DNA, a process dubbed DNA interference [6-10]. However, the natural functions and targets of DNA interference are poorly understood, and the mechanisms of DNA guide generation and target discrimination remain unknown. Here, we studied the *in vivo* activities of a bacterial Argonaute nuclease from *Clostridium butyricum* (CbAgo). We demonstrated that CbAgo targets multicopy genetic elements and suppresses propagation of plasmids and infection by phages. CbAgo induces DNA interference between homologous sequences and triggers DNA degradation at double-strand breaks in the target DNA. Loading of CbAgo with locus-specific small DNA guides depends on both its intrinsic endonuclease activity and the cellular double-strand break repair machinery. A similar interplay was reported for acquisition of new spacers during CRISPR adaptation, and prokaryotic genomes encoding pAgo nucleases are enriched in CRISPR-Cas systems. These results identify molecular mechanisms that generate guides for DNA interference and suggest common principles of recognition of foreign nucleic acids by prokaryotic defense systems.

Argonaute (Ago) proteins are a ubiquitous family of guide-dependent nucleases found in all three domains of life [3,5,11,12]. Prokaryotic Agos (pAgos) are extremely diverse in comparison with eukaryotic Agos and are disseminated among bacterial and archaeal lineages by horizontal gene transfer [3-5]. In contrast to their eukaryotic counterparts, several studied pAgos were shown to act *in vitro* as nucleases with distinct specificity toward DNA targets [6-8,10,13-17]. The majority of studied pAgos also use short single-stranded DNA molecules as guides for their endonuclease activity, but the molecular mechanism of DNA guide generation in bacterial cells has remained poorly understood. pAgos were proposed to protect bacteria against foreign DNA [9,18], but the ability of pAgos to fight genuine invaders such as phages was not demonstrated and the factors that could instruct pAgos for the recognition of foreign genetic elements have remained unknown.

Here, we investigate the mechanism of DNA interference by a recently characterized pAgo nuclease, CbAgo from a mesophilic bacterium *Clostridium butyricum*, which acts as an efficient DNA-guided DNA nuclease under ambient conditions *in vitro* [6,7]. We demonstrate that in bacterial cells CbAgo preferentially targets multicopy genetic elements and induces DNA interference between homologous sequences, acting in cooperation with the cellular DNA repair machinery.

CbAgo binds small DNAs from *ter* sites

To study the cellular activities of CbAgo, we expressed and purified it from *Escherichia coli* cells and analyzed associated nucleic acids (Fig. 1a). CbAgo was bound to small guide DNAs (smDNAs) of 14-23 nt, which are highly diverse and have a moderate AT-bias near the 5'-end and downstream of the site of target cleavage, suggesting that their biogenesis might depend on melting of these DNA regions in the context of double-stranded DNA (Extended data Fig. 1).

Chromosomal mapping of CbAgo-bound smDNAs revealed that they are distributed through the whole genome, with two large peaks at the region of replication termination (Fig. 1b,c). The smDNA hotspots are bounded by the replication termination sites (*ter* sites) from the outside and are asymmetric, with the higher peak corresponding to *terC* and the lower to *terA*. This correlates with the lengths of the right and left replichores and different frequencies of replication termination at *terC* and *terA* [19]. No smDNA peaks at *ter* sites were observed in a strain lacking Tus, the protein that binds *ter* sites and limits replisome progression. Instead, a single peak was located exactly opposite the origin of replication, close to *terC*, likely corresponding to the main site of replication termination in this strain (Extended Data Fig. 2b). Therefore, replication termination by Tus-DNA complexes leads to preferential smDNA generation at *ter* sites.

DNA interference between multicopy elements

When CbAgo was expressed from a plasmid with a region of homology to the chromosome, the *araC* gene, an additional strong peak of smDNAs appeared around the

genomic *ara* locus (30-40 kb in each direction) (Fig. 1b,d). In a strain containing a plasmid with the *lacI* gene (with chromosomal CbAgo in this case) two smDNA peaks were observed around two chromosomal *lacI* loci present in this strain (Extended data Fig. 2c,d). The presence of plasmids therefore directs CbAgo toward homologous chromosomal regions and induces generation of smDNA guides from flanking sequences.

The targeting of homologous plasmid and chromosomal loci by CbAgo suggests that the presence of multicopy elements in the chromosome may also trigger generation of smDNA guides. Indeed, we revealed peaks of smDNAs around multicopy genomic sequences, primarily ribosomal RNA (rRNA) operons and IS elements (Fig. 1e). The peak width in each case was in the range of dozen kilobases, far beyond the borders of ribosomal DNA (rDNA) and IS elements. These results demonstrate that CbAgo targets multicopy sequences of both plasmid and genomic origin and triggers processing of smDNA guides from homologous regions.

Role of the catalytic activity of CbAgo

To study the role of the nuclease activity of CbAgo in generation of smDNA guides, we analyzed a catalytically impaired CbAgo variant, dCbAgo (catalytically dead CbAgo) that contains substitutions of two out of four catalytic residues in the active site and is incapable of cleaving DNA *in vitro*[7]. dCbAgo was loaded with smDNAs *in vivo* suggesting that its endonuclease activity is not strictly required to generate smDNA guides (Extended Data Fig. 1a), but their profile along the genome was changed significantly. Preferential generation of smDNAs at multicopy sequences – *araC*, rDNA and IS elements – was completely abolished in the dCbAgo mutant (Fig. 1d,e). Analysis of the ratio between smDNAs associated with wild-type CbAgo and dCbAgo allowed precise mapping of the chromosomal regions whose targeting depended on the catalytic activity of CbAgo (lower panels in Fig. 1c-e; Extended Data Fig. 2i). This analysis confirmed the key role of CbAgo in smDNA biogenesis at the rDNA loci and IS elements, especially the multicopy families of IS1 and IS3. In particular, all copies of IS1 were found in the areas with the WT/dCbAgo ratio of >1, a dramatic enrichment in comparison with random sampling ($p<0.0001$). In contrast, smDNA peaks in the *ter* region were still observed for dCbAgo, although they

were reduced compared to the wild-type protein (Fig. 1c; Extended data Fig. 2e). Therefore, the catalytic activity of CbAgo is indispensable for DNA interference between multicopy sequences and contributes to generation of smDNAs at the termination sites.

Asymmetry in smDNA processing

Inspection of smDNA distribution around the genomic *ara* and *lac* loci (undergoing DNA interference directed by the plasmid-encoded genes) or in rDNA loci (undergoing intrachromosomal DNA interference) revealed that the majority of smDNA reads corresponded to the DNA strand whose 3'-teminus was oriented towards the target gene (Fig. 2a,b, WT CbAgo; Extended Data Fig. 3a,b). In the *ter* region, most smDNAs were produced from the 3'-terminated DNA strand at each *ter* site (Fig. 2c). The outer boundaries of smDNA peaks at *araC, lacI* and rDNA loci were defined by Chi (χ) sites oriented in the 5'→3'-direction in the preferentially targeted DNA strand (Fig. 2a,b; Extended Data Fig. 3a,b). The inner boundaries of the *terA* and *terC* peaks of smDNAs corresponded to the first Chi site located in the 3'-terminated strands before the respective *ter* sites (Fig. 2c). SmDNA distribution around Chi sites across a 1 megabase region lacking strong smDNA peaks also demonstrated enrichment for smDNAs coming from the 3'-terminated strand at the 3'-side of Chi sequences, for Chi sites from both genomic strands (Extended Data Fig. 3c).

Chi sites are species-specific motifs recognized by the RecBCD or closely related AddAB helicase-nucleases, which participate in double-strand break (DSB) repair and homologous recombination [20-22]. The observed asymmetry of smDNAs and their dependence on the position of Chi sites suggests that smDNA are generated from the 3'-terminated strands at the sites of DSBs with participation of RecBCD. In the case of *araC*, *lacI* and multicopy elements, smDNAs are produced from both ends of the DSB presumably formed at the target loci. In the case of *ter* sites, smDNAs are likely produced from double-stranded DNA ends, formed at *terA* and *terC*, by RecBCD moving in reverse direction relative to prior movement of the arrested replication forks.

Role of the DSB repair machinery

To explore the role of the DSB repair machinery in smDNA biogenesis, we analyzed strains with knockouts of individual components of RecBCD (*recBrecD* and *recC*) and of RecA. The amount of CbAgo-bound smDNAs did not change in the *recBrecD* and *recC* mutants (Fig. 2d), but the polarity for smDNAs around Chi sites was eliminated on the whole genome level in both strains, confirming that it is defined by RecBCD (Extended Data Fig. 3c).[23]

Double *recBrecD* knockout eliminated the enrichment of smDNAs at *araC*, rDNA and *terA*/*terC* loci indicating that RecBCD orchestrates DNA processing at multicopy sequences and the replication termination sites (Fig. 2b,c; Extended data Fig. 2f and 3b). Instead of two smDNA peaks at *terA* and *terC*, a single peak that spread beyond the *ter* sites and was independent of Chi sites was present in the mutant cells. This is likely due to generation of DSBs in the terminus by cell division with incomplete replication, previously observed for *recB* or *recC* mutants [24,25], followed by smDNA processing by CbAgo and/or other cellular nucleases. The *recC* knockout similarly eliminated peaks at *ter* sites (Fig. 2c). In contrast, the peaks at *araC* and rDNA loci were still present in the *recC* strain, though they became independent of Chi sites and symmetrical relative to the two DNA strands (Fig. 2b, Extended data Fig. 2g and 3b). This indicates that the helicase-nuclease activities of RecB or RecD (which form no functional complex in the absence of RecC [23]) may have independent roles in DNA interference, while RecC makes DNA processing dependent on Chi sites through the RecBCD assembly. Together, these results reveal an essential role of the RecBCD complex in biogenesis of smDNA guides from both multicopy sequences and the replication termination sites.

The amount of smDNAs loaded into CbAgo was strongly increased in the *recA* strain (Fig. 2d). This was accompanied by loss of specific enrichment at *ter*, *araC* or multicopy loci (Fig. 2b,c and Extended Data Fig. 2h). SmDNA distribution on the whole-genome level showed strong dependence on the Chi sites suggesting the involvement of RecBCD (Extended Data Fig. 3c). Thus, RecA guards the genome from excessive processing by

CbAgo and RecBCD, a phenomenon known as 'reckless' DNA degradation due to persistence of DNA ends and their continuous degradation by RecBCD in *recA* strains[26].

Targeting of chromosomal DNA by CbAgo might be toxic for cells as in the case of CRISPR-Cas systems, which avoid autoimmunity by requiring protospacer-associated motifs in the target DNA for self-nonself discrimination during interference [27,28]. We found only small differences in the kinetics of cell growth for wild-type or *tus* strains with and without expression of CbAgo. In contrast, the growth in the presence of CbAgo was slower in strains with defects in the DSB repair system (*recBrecD* and *recA*) (Extended data Fig. 4). This suggested that in wild-type cells CbAgo-induced DSBs are efficiently repaired. In agreement with this, whole-genome sequencing did not reveal any significant DNA degradation at the sites of preferential smDNA processing (*araC*, *ter* or rDNA) upon CbAgo expression (Extended data Fig. 5).

Analysis of the ratio of smDNAs originating from the two genomic strands on the whole-genome scale revealed a small but significant bias toward DNA strands whose 3'-termini were oriented opposite to the direction of each replichore, indicating that smDNAs may preferentially originate from the lagging DNA strand during replication (Extended Data Fig. 3d). This additional asymmetry was dependent on the catalytic activity of CbAgo but not on RecBCD, suggesting that CbAgo can target the discontinuous lagging DNA strand during replication independently of RecBCD.

DSB processing by CbAgo

The dependence of smDNA production on Chi sites and RecBCD implies the formation of DSBs in the target regions during CbAgo-induced DNA interference. To directly test the role of DSBs in smDNA biogenesis, we analyzed CbAgo-associated smDNAs in *E. coli* strains with engineered DSBs, induced in the *lac* locus by either expression of the I-SceI meganuclease, recognizing its respective site in the genome [29], or by a long palindrome (*pal*) processed by the host SbcCD (Mre11-Rad50) complex [30] (Fig. 2e). In both cases, we observed highly efficient loading of smDNAs from the sites of breaks into CbAgo, with the peak size greatly exceeding the peaks at the *ter* sites (Fig. 2f and Extended Data Fig. 6a). The number of smDNAs was smaller in a strain with a mutated I-SceI site, which is

cleaved less efficiently[29], demonstrating that smDNA production depends on the efficiency of DSB formation (Extended Data Fig. 6a). Inactive dCbAgo was also loaded with smDNAs at DSBs, although with a somewhat lower efficiency than wild-type CbAgo (Extended Data Fig. 6a), suggesting that the catalytic activity of CbAgo is not strictly required for generation of smDNA guides from induced DSBs. Processing of engineered DSB involves RecBCD since the boundaries of the smDNA peaks are defined by Chi sites, with more smDNAs being produced from the 3'-terminated strand at each end of the DSB (Fig. 2f, Extended Data Fig. 6a).

To analyze the integrity of genomic DNA around engineered DSBs, we employed high-throughput sequencing. As expected, in wild-type cells the DNA content gradually decreased towards termination sites (Fig. 2g, -DSB/+Cb). No obvious decrease in the genomic DNA content was observed at the site of palindrome-induced DSB in the absence of CbAgo indicating that it is efficiently repaired (Fig. 2g, +DSB/-Cb, and Ref. [31]). In contrast, expression of wild-type but not catalytically impaired CbAgo strongly decreased DNA coverage at the site of engineered DSB (as far as 330 kb around the original DSB site) (Fig. 2g, +DSB/+Cb; Extended data Fig. 6b,c). In the case of the permanent DSB introduced by the I-SceI expression, expression of CbAgo further stimulated DNA degradation in this region which could already be detected in its absence (Extended Data Fig. 6d,e). Therefore, CbAgo activity triggers substantial DNA loss at genomic regions flanking DSBs.

CbAgo eliminates plasmids

Our results indicate that smDNA guides loaded into CbAgo are preferentially generated from multicopy sequences and the sites of DSBs. These observations suggest that CbAgo may target mobile genetic elements, such as transposons, plasmids and phages, which have multicopy nature and form free DNA ends in their life cycle. Indeed, a disproportionally large fraction of smDNAs bound to CbAgo (up to 20%) were derived from plasmid DNA, one-order of magnitude higher than expected after accounting for the plasmid copy numbers and relative replicon lengths (Fig. 3a). A similar enrichment was observed independently of whether CbAgo was expressed from plasmid or chromosome

and whether the plasmid had any homology to chromosomal DNA or not. SmDNAs evenly mapped to both plasmid strands (Extended data Fig. 7a,b). Consistently, analysis of plasmid DNA content by high-throughput sequencing did not reveal specific regions of preferential DNA degradation in the presence of CbAgo (Extended data Fig. 7d).

To reveal whether RecBCD participates in plasmid processing by CbAgo, we compared two plasmids with identical sequences, except for four tandem Chi sites present in one of them. The smDNA distribution had a clear bias for smDNAs being produced from the 3'-side of the Chi sites, showing the involvement of RecBCD (Fig. 3b). We further analyzed CbAgo-bound smDNAs isolated from cells with a much larger foreign replicon, a single-copy F' plasmid containing a large genomic insert (~100 kb) in the conjugative F factor. CbAgo guides were >10-fold enriched with sequences derived from the F' plasmid. Most smDNAs were generated from the F factor part lacking Chi sites while their numbers were greatly decreased in the chromosomal insert containing multiple Chi sites (Extended data Fig. 7e-g). Thus, the presence of Chi sites contributes to discrimination between chromosomal and foreign DNA and allows preferential generation of smDNA guides from plasmid sequences.

We then tested the effect of CbAgo expression on plasmid maintenance in bacterial cells. Control strains lacking CbAgo or containing inactive CbAgo fully maintained the plasmids (Fig. 3c). Remarkably, expression of wild-type CbAgo led to fast and quantitative loss of plasmids from various incompatibility groups (Fig. 3c and Extended data Fig. 8). Taken together, these results indicate that CbAgo is strongly enriched in plasmid-derived guides and employs them to suppress plasmid propagation.

CbAgo provides protection against phages

To explore the ability of CbAgo to protect *E. coli* from bacteriophages, we analyzed infection with phages P1, T7 and M13, which have different genome organization and infection cycles. CbAgo did not protect bacteria from T7, a lytic phage with a linear dsDNA genome (Supplementary Table 1). In contrast, CbAgo strongly decreased the titers of M13, a chronic phage with a circular ssDNA genome (15 to 270-fold at 4 to 8 hours post infection) (Fig. 3d). CbAgo-bound smDNAs contained M13-derived sequences, which

mapped not only to the genomic strand of phage DNA but also to the complementary strand, indicating that CbAgo is loaded with smDNAs during phage replication (Extended data Fig. 7c).

CbAgo was even more efficient in protecting the cells against infection with P1, a lytic phage (P1vir) with a circularly permutated dsDNA genome. Infection of *E. coli* cells with P1 even at low multiplicity of infection (MOI, 0.1 phages per bacterium) led to complete lysis of the bacterial culture after 8 hours. In contrast, expression of CbAgo delayed bacterial lysis in liquid cultures at high MOI (5), when most cells get infected, supported steady culture growth at intermediate MOI (1), and almost fully protected the cells at low MOI (0.1) (Fig. 3e, Extended data Fig. 9a). Consistently, P1 titers dropped dramatically (up to three-four orders of magnitude) in the presence of CbAgo at different times of infection (Fig. 3d and Extended data Fig. 9a). CbAgo also impaired formation of P1 plaques in an independent plaque assay (inset in Fig. 3e). Inactive dCbAgo had almost no effect on phage titers or cell lysis (except for initial stages of infection at low MOI, Extended data Fig. 9a,b), indicating that the endonuclease activity of CbAgo is essential for protection. In contrast, dCbAgo had a smaller than wild-type CbAgo but still significant effect on infection with phage M13 (Fig. 3d). It is plausible that dCbAgo might provide modest protection against M13 by guide-dependent binding to its single-stranded DNA even without cleavage. Notably, many pAgos lack endonuclease activity due to natural substitutions in their active sites[3-5], suggesting that cleavage of target DNA is not essential for their function.

Discussion

Our results show that CbAgo induces DNA interference between homologous/multicopy sequences, targets DNA regions containing breaks and free ends, and relies on the cellular DNA break repair machinery for smDNA loading. Since pAgo proteins are spread by horizontal gene transfer and their phylogeny does not correspond to the phylogeny of host species [3-5], they may have adapted to cooperate with various types of DNA processing and recombination machineries. Analysis of co-occurrence of pAgos with cellular DSB repair systems indeed shows diverse combinations in various bacterial

species, with a substantial fraction of catalytically active pAgos found in the same genomes with AddAB or RecBCD (47% of fully sequenced genomes containing active pAgos) (Fig. 4a, Extended Data Fig. 10a). Prokaryotes have another defense system, CRISPR-Cas, whose activity is also based on complementary recognition of target genetic elements, but that uses different components and biochemistry. Spacer acquisition by type I and II CRISPR-Cas systems was also shown to be dependent on RecBCD/AddAB and on DNA replication resulting in specific targeting of the *ter* region and DSBs [32-34]. Many phages, such as T7 or λ, encode inhibitors of RecBCD which may thus help them to cope with multiple defense systems in host bacteria [35]. Remarkably, species with active pAgos have a substantially higher content of CRISPR-Cas systems in their genomes (77% of genomes) than prokaryotes with inactive pAgos or without pAgos (32-40%) (Fig. 4a,b, Extended data Fig. 10a). Furthermore, 47% of genomes with active pAgos contain two or more CRISPR-Cas systems, in contrast to only 6-10% for other genomes (Fig. 4b). Among CRISPR-Cas types, type I (subtypes I-A, I-B, I-D) and type III (all subtypes) systems are strongly enriched in the genomes with pAgos, while type II is somewhat depleted (Fig. 4d and Extended data Fig. 10b). *C. butyricum* contains a type I-B CRISPR-Cas system (located on a chromid) together with CbAgo (located on the chromosome). The ability of pAgos to target DSBs and their co-occurrence with CRISPR-Cas suggests that the two systems may cooperate in foreign DNA targeting, which deserves further investigation.

We demonstrate that CbAgo from *C. butyricum* can function in anti-plasmid and anti-phage defense in *E. coli*. Similarly, other bacterial defense systems were shown to be active in heterologous hosts [36,37]. The finding that pAgo uses DNA guides to protect bacteria against invaders suggests a conserved function of Ago-mediated nucleic acid interference (RNAi and DNAi) as an ancient defense system that has survived billions years of evolution in both prokaryotes and eukaryotes [38]. The functional activity of CbAgo against invading genetic elements, plasmids and phages, can potentially be explained by a combination of several mechanisms (Fig. 4d): (i) the higher copy number of invader DNA allowing the loading of multiple guides into pAgo and resulting in more efficient DNA targeting over unique genomic regions; (ii) the ability of pAgo to target DSBs and replication intermediates, which can form in multiple copies of plasmids and phages; such small replicons can then be completely degraded by cellular nucleases; (iii) the relative

absence of Chi sites in foreign DNA that makes it a better substrate for RecBCD/AddAB processing [39]; this allows efficient targeting of even low copy number replicons, as illustrated by differential generation of smDNA guides from the two portions of the F' plasmid with different densities of Chi sites. Loading of locus-specific guides into pAgo increases the concentration of effector complexes and likely induces DNA interference leading to further processing of new smDNAs from the target locus, amplification of guide DNAs and effective destruction of the target (Fig. 4d). The role of the proposed mechanism in anti-phage defense by various pAgos should be confirmed in future studies. CbAgo also targets specific genomic regions, including multicopy loci, the *ter* domain of the chromosome and the sites of DSBs. CbAgo-dependent DNA processing may possibly play roles in various genomic processes such as elimination of repetitive genetic elements, DNA repair and recombination, and suicidal response [12]. pAgos might also participate in the resolution of DNA intermediates during replication termination, as was recently proposed for another active pAgo nuclease, TtAgo from *Thermus thermophilus* [40]. Specific targeting of DSBs by pAgos can potentially be used for genomics applications, in particular as an instrument to study the genome architecture and DNA processing in both prokaryotic and eukaryotic systems.

References

1. Gebert, D. & Rosenkranz, D. RNA-based regulation of transposon expression. *Wiley interdisciplinary reviews. RNA* **6**, 687-708, (2015).
2. Meister, G. Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet* **14**, 447-459, (2013).
3. Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology direct* **4**, 29, (2009).
4. Ryazansky, S., Kulbachinskiy, A. & Aravin, A. A. The Expanded Universe of Prokaryotic Argonaute Proteins. *mBio* **9**, e01935-01918, (2018).
5. Swarts, D. C. et al. The evolutionary journey of Argonaute proteins. *Nature structural & molecular biology* **21**, 743-753, (2014).
6. Hegge, J. W. et al. DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute. *Nucleic acids research* **47**, 5809-5821, (2019).
7. Kuzmenko, A., Yudin, D., Ryazansky, S., Kulbachinskiy, A. & Aravin, A. A. Programmable DNA cleavage by Ago nucleases from mesophilic bacteria

Clostridium butyricum and Limnothrix rosea. *Nucleic acids research* **47**, 5822-5836, (2019).
8. Swarts, D. C. et al. Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA. *Nucleic acids research* **43**, 5120-5129, (2015).
9. Swarts, D. C. et al. DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**, 258-261, (2014).
10. Zander, A. et al. Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii. *Nature microbiology* **2**, 17034, (2017).
11. Willkomm, S., Makarova, K. & Grohmann, D. DNA-silencing by prokaryotic Argonaute proteins adds a new layer of defence against invading nucleic acids. *FEMS microbiology reviews* **42**, 376-387, (2018).
12. Lisitskaya, L., Aravin, A. A. & Kulbachinskiy, A. DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. *Nature communications* **9**, 5165, (2018).
13. Kaya, E. et al. A bacterial Argonaute with noncanonical guide RNA specificity. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 4057-4062, (2016).
14. Sheng, G. et al. Structure-based cleavage mechanism of Thermus thermophilus Argonaute DNA guide strand-mediated DNA target cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 652-657, (2014).
15. Willkomm, S. et al. Structural and mechanistic insights into an archaeal DNA-guided Argonaute protein. *Nature microbiology* **2**, 17035, (2017).
16. Olina, A. et al. Genome-wide DNA sampling by Ago nuclease from the cyanobacterium Synechococcus elongatus. *RNA biology*, 1-12, (2020).
17. Swarts, D. C. et al. Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. *Molecular cell* **65**, 985-998, (2017).
18. Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. K. & Aravin, A. A. Bacterial argonaute samples the transcriptome to identify foreign DNA. *Molecular cell* **51**, 594-605, (2013).
19. Duggin, I. G. & Bell, S. D. Termination structures in the Escherichia coli chromosome replication fork trap. *Journal of molecular biology* **387**, 532-539, (2009).
20. Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiology and molecular biology reviews : MMBR* **72**, 642-671, Table of Contents, (2008).
21. Smith, G. R. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiology and molecular biology reviews : MMBR* **76**, 217-228, (2012).
22. Wigley, D. B. Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nature reviews. Microbiology* **11**, 9-13, (2013).
23. Chaudhury, A. M. & Smith, G. R. Escherichia coli recBC deletion mutants. *Journal of bacteriology* **160**, 788-791, (1984).

24. Sinha, A. K. et al. Division-induced DNA double strand breaks in the chromosome terminus region of Escherichia coli lacking RecBCD DNA repair enzyme. *PLoS genetics* **13**, e1006895, (2017).
25. Sinha, A. K. et al. Broken replication forks trigger heritable DNA breaks in the terminus of a circular chromosome. *PLoS genetics* **14**, e1007256, (2018).
26. Capaldo, F. N. & Barbour, S. D. DNA content, synthesis and integrity in dividing and non-dividing cells of rec- strains of Escherichia coli K12. *Journal of molecular biology* **91**, 53-66, (1975).
27. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568-571, (2010).
28. Westra, E. R. et al. Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS genetics* **9**, e1003742, (2013).
29. White, M. A., Azeroglu, B., Lopez-Vernaza, M. A., Hasan, A. M. M. & Leach, D. R. F. RecBCD coordinates repair of two ends at a DNA double-strand break, preventing aberrant chromosome amplification. *Nucleic acids research* **46**, 6670-6682, (2018).
30. Eykelenboom, J. K., Blackwood, J. K., Okely, E. & Leach, D. R. SbcCD causes a double-strand break at a DNA palindrome in the Escherichia coli chromosome. *Molecular cell* **29**, 644-651, (2008).
31. White, M. A., Darmon, E., Lopez-Vernaza, M. A. & Leach, D. R. F. DNA double strand break repair in Escherichia coli perturbs cell division and chromosome dynamics. *PLoS genetics* **16**, e1008473, (2020).
32. Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544**, 101-104, (2017).
33. Levy, A. et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510, (2015).
34. Ivancic-Bace, I., Cass, S. D., Wearne, S. J. & Bolt, E. L. Different genome stability proteins underpin primed and naive adaptation in E. coli CRISPR-Cas immunity. *Nucleic acids research* **43**, 10821-10830, (2015).
35. Bobay, L. M., Touchon, M. & Rocha, E. P. Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. *PLoS genetics* **9**, e1003825, (2013).
36. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, (2018).
37. Ofir, G. et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nature microbiology* **3**, 90-98, (2018).
38. Koonin, E. V. Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biology direct* **12**, 5, (2017).
39. El Karoui, M., Biaudet, V., Schbath, S. & Gruss, A. Characteristics of Chi distribution on different bacterial genomes. *Research in microbiology* **150**, 579-587, (1999).
40. Jolly, S. M. et al. A DNA-guided Argonaute Protein Functions in DNA Replication in Thermus thermophilus. *bioRxiv*, 869172, (2019).

Fig. 1. CbAgo targets specific genomic regions.

(a) CbAgo is expressed from the plasmid or chromosome, followed by isolation of CbAgo-associated smDNAs and their sequencing. (b) Distribution of smDNA guides associated with CbAgo along the bacterial chromosome. Replication origin and termination sites and the directions of replichores are indicated. (c) SmDNA peaks in the *ter* region. *TerA* and *terC* are the innermost replication termination sites bound by the Tus protein. (d) SmDNA peak around the *araC* locus. (e) SmDNA profiles in a chromosomal region containing four rRNA operons (blue) and a cluster of IS elements (black). WT CbAgo, top; dCbAgo (dCb), middle; the ratio of smDNAs (WT to dCb), bottom. SmDNA profiles are shown in RPKM (reads per kilobase per million).

Fig. 2. CbAgo cooperates with RecBCD and targets DSBs.

(a) The RecBCD complex (RecB, 3'-5' helicase and nuclease; RecD, 5'-3' helicase; RecC, Chi site recognition) recognizes Chi in the 3'-terminated strand (red) during DSB processing. This switches its activity from preferential cleavage of the 3'-terminated strand to preferential cleavage of the 5'-terminated strand (green) thus generating a 3'-terminated single strand for RecA loading. (b) and (c) Strand-specific distribution of smDNAs in the *araC* locus (b) and *ter* region (c) for strains with various genetic backgrounds. Reads from the plus and minus genomic strands are shown in green and red, respectively. Positions of Chi sites in surrounding genomic regions are indicated (forward for the plus strand and reverse for the minus strand); the closest Chi sites in the correct orientation are shown with dotted lines. LR and RR, left and right replichores. (d) SmDNA loading into CbAgo in various bacterial strains (CbAgo expression was analyzed by Western blotting, WB). For gel source data, see Supplementary Figure 1. (e) Formation of DSBs at the I-SceI site or a 246 bp palindrome inserted in the *lacZ* locus. (f) Peaks of CbAgo-bound smDNAs around engineered DSBs (palindrome or I-SceI). (g) Genomic DNA coverage at the site of palindrome-induced DSB in strains without palindrome with expression of CbAgo (left), with palindrome without CbAgo (middle), and with both palindrome and CbAgo (right). CbAgo triggers DNA loss around the DSB with overreplication of genomic DNA at the site of termination.

Fig. 3. CbAgo interferes with plasmids and phage infection.

(a) Enrichment of CbAgo with plasmid-derived smDNAs. The observed (blue) and expected (green) read numbers, based on random sampling of the genomic and plasmid DNA with normalization to the plasmid copy number (pCbAgo, plasmid-encoded CbAgo; gCbAgo, genomic CbAgo). pNonChi and pChi are pSRK derivatives without regions of homology to the genome; pChi contains four tandem Chi sites; pBAD and pET28 contain *araC* and *lacI*, respectively. (b) Asymmetry in smDNA distribution around the plasmid Chi sites (the ratio of smDNAs from the plus plasmid strand, containing Chi in the rightward orientation, for pChi and pNonChi). (c) CbAgo-induced loss of plasmids in strains with genomic CbAgo. The cells were grown for 5-9 passages in the absence of selection for plasmid genes, to allow elimination of plasmids. Means and standard deviations from at least three independent measurements. (d) Titers of P1 (left, MOI=0.1) and M13 (right) at

different times of infection depending on the expression of chromosomal CbAgo or dCbAgo. PFU, plaque forming units. Means and standard deviations from three-four independent measurements. Statistical analysis was performed using the Scheffe's test for multiple comparison of means after normalization of data by log-transformation. (*p<0.05, **p<0.01, ***p<0.001). (e) Bacterial culture growth during P1 infection with different MOI in strain with wild-type CbAgo. Means and standard deviations from three independent experiments. *Inset*, plating of P1 with identical titers on the strains lacking CbAgo or expressing wild-type CbAgo or dCbAgo.

Fig. 4. Co-occurrence of pAgos, DSB repair systems and CRISPR-Cas, and the mechanism of DNA interference.

(a) Phylogenetic tree of three major groups of pAgos ('long-A', 'long-B' and 'short'[4]). The vast majority (94%) of long-A pAgos contain a predicted nuclease site, while all long-B and short pAgo are catalytically inactive (inner circle). The middle and outer circles show co-occurrence of pAgos with RecBCD/AddAB and CRISPR-Cas. (b) The numbers of CRISPR/Cas systems in the fully assembled genomes of species encoding pAgos from different groups. (c) The distribution of various types of CRISPR-Cas systems (indicated below the plot) in the genomes with pAgos. The number of genomes containing pAgos from each group is indicated. (d) Proposed mechanism of DNA interference by CbAgo. CbAgo targets invader DNA and chromosomal loci with regions of homology to plasmids (red) and multicopy genomic elements (violet).

Methods

Plasmids and strains

Plasmids and bacterial strains used in this study are listed in Supplementary Tables 2 and S3. All strains are isogenic to either *E. coli* BL21(DE3), MG1655 or BW27784. *E. coli* cells were routinely cultivated in standard LB Miller broth (2% tryptone, 0.5% yeast extract, 1% NaCl, pH 7.0) with the addition of appropriate antibiotics (ampicillin, 100 μg/ml; kanamycin, 50 μg/ml; chloramphenicol, 12.5 μg/ml). All plasmids were introduced into recipient strains using a standard electroporation protocol according to the manufacturer's instructions (BioRad Gene Pulser Xcell, 2.5 kV, 0.2 cm cuvettes).

For the construction of the pChi and pNonChi plasmids, a tandem of four consecutive Chi-sites was inserted into the pSRKKm plasmid via the Gibson assembly technique. Briefly, 5 μM of oligonucleotides 4xChi_fwd and 4xChi_rev (or 4xnonChi_fwd and 4xnonChi_rev) (Supplementary Table 4) were mixed together in 50 μl of 1^x annealing buffer (10 mM Tris-HCl, 50 mM NaCl, 1 mM EDTA, pH 8.0), incubated at 98 °C for 5 minutes, slowly cooled down to 25 °C in a thermocycler, and cloned into pSRKKm linearized at the NheI and EcoRI sites, in order to remove a 1661 bp fragment containing regions of homology with the *E. coli* chromosome (*lacI*, *lacZ*α).

E. coli knock-out strains were obtained via Red-mediated gene disruption method essentially as described in [41]. Briefly, kanamycin- or chloramphenicol-resistance cassettes were amplified with primers that contain 45 nt homology arms to the genomic regions flanking the target ORF from pKD4 or pKD3 plasmids, respectively, gel-purified and transformed into electrocompetent BL21(DE3) cells expressing the lambda-Red recombinase system under the control of the araBAD inducible promoter from the pKD46 plasmid. After 4 hours of recovery, the cells were plated onto the selective LB plates. The mutant alleles with selective markers were then transferred to parent strains by P1 transduction. The correct genomic integration events were verified by PCR and sequencing. The absence of the expression of the B, C and D subunits of the RecBCD complex was additionally checked by Western blotting with anti-RecB, anti-RecD and anti-RecC rabbit antisera (kindly provided by Dr. Gerry Smith).

E. coli strains with chromosomal insertions of the CbAgo gene were obtained starting from the MG1655Z1 strain containing a Z1 cassette with spectinomycin resistance (*LacI, TetR, SpR*) in the chromosomal *attB* site [42]. MG1655Z1::CbAgo was constructed by insertion of a PCR product disrupting the Sp^R gene inside the Z1 cassette and containing two strong transcription terminators (t0 and BBaT1006), N-His6 CbAgo under the control of TetP promoter, BBaT1001 terminator and the *cat* gene. Positive Sp^S Cm^R clones were selected and verified by PCR and sequencing. This strain was used as a donor strain for P1 transduction of the wild-type MG1655Z1, NEB Turbo and strains containing the palindrome and I-SceI cleavage sites inside *lacZ*. The I-SceI cleavage site (TAGGGATAACAGGGTAAT) was present in the *lacZ* locus of the chromosome in the *E. coli* strain DL2917 [29]. The mutant variant of the cleavage site $I\text{-}SceI_{mut}$ in strain DL4977 contained two single-nucleotide substitutions relative to the wild-type sequence (TTGGGATAACAGGGTAAA, underlined). Strain DL2859 contained a 246 bp palindrome in *lacZ* and strain DL1777 was used as a control with the same genetic background [30]. Strains carrying dCbAgo were constructed using the same protocol.

Culture conditions for CbAgo expression and smDNA library preparation

In the case of plasmid-encoded CbAgo, the cells were transformed with plasmids encoding wild-type CbAgo or dCbAgo (pBAD/His B_CbAgo and pBAD/His B_dCbAgo, Supplementary Table 2). After overnight growth on LB plates, a single colony was transferred into 5 ml of LB media supplemented with appropriate antibiotics and 0.2% glucose and grown overnight at 37°C. The cells were inoculated into 0.5 L of fresh LB supplemented with 0.2% glucose and appropriate antibiotics and aerated at 37°C until OD_{600} = 0.3-0.4. At this point, the temperature was adjusted to 30°C and after 20 min CbAgo expression was induced by the addition of L-arabinose to 0.01% for 3 hours. In the case of genome-encoded CbAgo the cells were inoculated into LB medium that already contained the inducing agent (anhydrotetracycline, 200 ng/ml) and allowed to grow at 18°C until OD_{600} = 1.0. The cells were collected by centrifugation at 7000 g, 4°C for 15 minutes and kept frozen at -20°C.

Isolation of smDNAs

All purification steps were carried out at 4°C. The cell pellet was resuspended in lysis buffer (50 mM Tris-HCl, 250 mM NaCl, 5% glycerol, 0.5 mM β-mercaptoethanol, pH 7.4 at 23°C) supplemented with EDTA-free protease inhibitor cocktail (Roche) at approximately 7 ml per gram of wet weight and subjected to 3 rounds of disruption on a high-pressure homogenizer (Avestin) at 12000 psi. The cell lysate was clarified by centrifugation for 30 min at 35,000 g. CbAgo (dCbAgo) was pulled down on Talon metal affinity resin (Takara) charged with Co^{2+} for 2 hours at 4°C (300 μl of bead suspension per sample) with constant rocking. The beads were collected by centrifugation (300 g, 3 min) and washed one time with lysis buffer and three times with lysis buffer supplemented with 10 mM imidazole (140 ml in total). CbAgo-smDNA complexes were eluted in 500 μl of lysis buffer supplemented with 300 mM imidazole.

Nucleic acids associated with CbAgo were extracted with phenol:chloroform:isoamyl alcohol (PCI) mixture (25:24:1, pH 8.0, Applichem) according to a standard procedure. Briefly, the protein sample was mixed with an equal volume of PCI, vortexed intensively for 30 seconds and centrifuged at 13000 g for 3 min. The upper aqueous phase was carefully collected and treated twice with an equal volume of chloroform to eliminate any remaining traces of phenol. Nucleic acids were precipitated with three volumes of ethanol in the presence of a co-precipitating agent (PINK, Bioline) and dissolved in 20 μl of milliQ-grade water.

To determine the type of associated nucleic acids, the samples were treated with alkaline phosphatase (rSAP, NEB) at 37°C for 10 min with subsequent enzyme inactivation at 75°C for 7 min. The nucleic acids were subsequently radiolabeled with γ^{32}P-ATP using T4 polynucleotide kinase (NEB) according to the manufacturer's instructions. The samples were then divided into three equal parts and each part was treated with either DNase I (ThermoFisher), or RNase A (ThermoFisher) or left untreated. Radiolabeled nucleic acids were resolved on 19% denaturing polyacrylamide-urea gel and visualized on a Typhoon FLA 9500 imager (GE healthcare).

Western blotting

599 The level of CbAgo expression and the amount of CbAgo used for smDNA purification
600 was determined by Western blotting. Protein samples were mixed with 2x Laemmli sample
601 buffer (120 mM Tris-HCl, 4% SDS, 4% β-mercaptoethanol, 10% Glycerol, pH 6.8) and
602 heated at 95°C for 5 min, and then resolved by electrophoresis in a 4-20% Tris-glycine gel
603 (BioRad). Proteins were transferred onto a nitrocellulose membrane in Towbin transfer
604 buffer (25 mM Tris, 192 mM glycine, 20% methanol) using semi-dry procedure at 25 V, 1
605 A for 30 min (BioRad Trans-Blot Turbo). The transfer membrane was air-dried and then
606 washed in PBS (10 mM phosphate buffer, 137 mM NaCl, 2.7 mM KCl) for 5 min. The
607 membrane was blocked with blocking buffer (PBS, Tween-20 0.1% (v/v), non-fat milk 5%
608 (w/v) for 30 min at room temperature, and then incubated with anti-6xHis monoclonal
609 antibodies (1:1000, Sigma) for 1 h at room temperature. The membrane was washed 4
610 times with PBST buffer (PBS, Tween-20 0.1% (v/v)), and after that incubated with HRP-
611 conjugated anti-mouse secondary antibodies (1:10000, Sigma) for 1 h at room
612 temperature and washed again as described above. Antigen-antibody complexes were
613 detected with Immobilon ECL Ultra Western HRP substrate (Millipore) on a Chemidoc
614 XRS+ imager (BioRad).
615
616 **SmDNA library construction and sequencing**
617 Libraries for high throughput sequencing of smDNAs were prepared according to the
618 previously described bridged (splinted) ligation protocol [18]. For the purpose of
619 visualization, a 5 μl aliquot of the smDNA sample (1/4 of total volume) was
620 dephosphorylated and radiolabeled as described above. Unlabeled DNA was mixed with
621 the radiolabeled aliquot and 2^x urea sample buffer (8 M urea, 20 mM EDTA, 0.005%
622 Bromophenol Blue, 0.005% Xylene Cyanol) and resolved under denaturing conditions in a
623 15% polyacrylamide-urea gel. Gel slices corresponding to 14-25 nt smDNAs were crushed
624 and DNA was eluted in 0.4 M NaCl solution overnight at 21 °C with constant agitation.
625 After ethanol precipitation, DNA was resuspended in 20 μl of milliQ-grade water.
626 Illumina-compatible adaptors were ligated to both ends of smDNA. For this purpose, 8 μl
627 of 5^x Rapid Ligation buffer (ThermoFisher), 2 μl of 100 μM 5'-adaptor, 2 μl of 100 μM of 3'-
628 linker, 2 μl of 100 μM oligonucleotide bridge 1, 2 μl of 100 μM oligonucleotide bridge 2 and
629 2 units of T4 PNK (ThermoFisher) were added to 20 μl of the purified smDNA solution.

The reaction mixture was incubated overnight at room temperature. Ligated DNA fragments were recovered from 10% denaturing PAGE. The libraries were amplified and indexed according to the standard protocol (small RNA sequencing kit, NEB), except that PCR conditions were adjusted in order to prevent library overamplification. A series of test PCR were performed to optimize the amount of adaptor-ligated smDNA to achieve the desired amplification level in 3-4 cycles. smDNA libraries were sequenced using the HiSeq2500 platform (Illumina) in the rapid run mode (50 nt single-end reads). The list of all analyzed smDNA libraries is shown in Supplementary Table 5.

Genomic DNA library construction and sequencing

Genomic DNA was extracted from *E. coli* cells according to a published protocol [43]. For strains with chromosomal CbAgo genomic DNA was extracted from exponentially grown *E. coli* cells harvested at OD_{600}=1.0; for strains with plasmid-encoded CbAgo DNA was extracted from cells harvested at exponential (OD_{600}=0.5) and stationary (OD_{600}=6.0) phases. Preparation of genomic DNA libraries was carried out using the NEBNext Ultra II FS DNA Library Prep Kit (NEB) according to the manufacturer's instructions. The approximate insert size was selected to be in the range of 150-250 bp. Barcodes were introduced to both ends during library amplification step with NEBNext multiplex oligos for Illumina (NEB). Genomic DNA libraries were sequenced using HiSeq2500 platform (Illumina) in the rapid run mode (50 nt single end reads). The list of all analyzed genomic DNA libraries is shown in Supplementary Table 5.

Analysis of high-throughput sequencing data

All libraries were quality checked before further processing using FastQC (v. 0.11.8). The 3'-adaptor sequence (5'-TGGAATTCTCGGGTGCCAAGGC-3') was trimmed and reads with the length less than 14 nt were removed using cutadapt (v. 2.7). Reads were aligned onto the reference genomes (Refseq accession numbers: NC_012971.2 in the case of BL21(DE3) and its derivatives, NC_000913.3 in the case of MG1655 and its derivatives, and GSE107973 in the case of strains containing the I-SceI site) and corresponding plasmids (RefSeq CP014273.1 was used as a reference for the F'-episome) allowing zero mismatches via bowtie (v. 1.2.3, with parameters: -v 0 -m 1). Reads that failed to align due

to -m option were considered as multi-mappers and realigned using options: -a --best --strata -v 0 -m 10000. Multi-mappers aligned to both plasmid and chromosome loci were filtered out in order to avoid biases caused by different plasmid copy numbers. Genome arithmetic was done using bedtools (v. 2.28.0). SmDNA coverage was expressed in terms of RPKM (reads per kilobase per million) or RPM (reads per million) as indicated in the figure legends.

For the CbAgo/dCbAgo smDNA ratio (Fig. 1 and Extended Data Fig. 3), unique reads and multi-mappers were pooled together. For this purpose, read counts for each multi-mapper were normalized to the total number of locations it aligns to.

The chromosomal region lacking prominent hotspots of smDNA processing (between 2 and 3 Mb) was selected to create a metaplot of smDNA distribution around Chi sites (Extended data Fig. 3c). In this case, the number of reads that mapped to the DNA strand with the Chi site in the 5'-GCTGGTGG-3' orientation was calculated in 500 nt bins in the 20 Kb region centered around each Chi site and averaged.

To calculate the fraction of smDNAs generated at the sites of engineered DSBs, smDNAs mapped to the chromosomal region confined by Chi-sites nearest to the engineered DSB (357,321–489,948 bp in the case of I-SceI and 355,595–493,321 bp in the case of palindrome) were divided by the total number of chromosomal smDNAs in the library.

The expected number of smDNA reads mapped to a plasmid was calculated according to the random sampling model as follows: number of reads = (total number of reads)×(plasmid length)×(plasmid copy number)/((genome length) + (plasmid length)×(plasmid copy number)). Plasmid copy numbers were based on information available from the literature (12 for pBAD, 15 for pChi and pNonChi, 20 for pET28).

Nucleotide Logos (Extended Data Fig.1C, D) were generated using custom python scripts and WebLogo3 (v. 3.7). Only reads with the minimal length of 18 nt were taken into analysis and were all trimmed to 18 nt.

A permutation test was used to calculate the significance of the IS1 enrichment in the regions of preferential smDNA generation (Extended Data Fig. 2i). For this purpose, 10,000 samples of 29 IS1 elements randomly distributed across the *E. coli* genome were generated and intersected with the list of all genomic regions where the ratio of smDNA (CbAgo/dCbAgo) is greater than 1.

To analyze strand-specific distribution of smDNAs along the *E. coli* chromosome (Extended data Fig. 3d), the ratio between smDNAs mapped to the plus- and minus-strands in 1 Kb windows was calculated and plotted against the genomic coordinate. For the purpose of visualization, a rolling mean (10 Kb window, 1 Kb step) was added to the plots.

Genomic DNA libraries were processed essentially as described above. Genome coverage in 1 Kb windows was calculated, smoothed with Loess regression and plotted against the chromosome coordinate.

All plots were generated in R (v. 3.6.1) using custom scripts.

Determination of the efficiency of plasmid loss

For experiments on plasmid loss, *E. coli* MG1655Z1 strains containing chromosomal CbAgo/dCbAgo or lacking CbAgo (Supplementary Table 3) were transformed with plasmids from various incompatibility groups (Supplementary Table 2). Several colonies of transformed cell cultures from LB plates were inoculated into liquid LB medium with the addition of anhydrotetracycline (aTc) (0.1 μg/ml) and grown at 25°C until OD_{600}~0.5. Then, glycerol was added to the cell culture to the final concentration of 20%, 200 μl aliquots were frozen in liquid nitrogen and stored at -70 °C.

Aliquots of frozen cells of *E. coli* MG1655Z1::CbAgo, MG1655Z1::dCbAgo or control MG1655Z1 were thawed on ice and inoculated into liquid LB with the addition of aTc and antibiotic corresponding to the plasmid used (Supplementary Table 2). The cell culture was grown at 25 °C and passaged in LB with the addition of aTc and without any antibiotics every 12 hours. The percentage of antibiotic resistant cells containing the plasmid was determined by plating the cultures on agar media with and without antibiotic after 5 (RSF1010 and pSRKTc), 8 (pET28b), and 9 (pBR325 and pACYC) passages in Fig. 3, or after indicated number of passages in Extended data Fig. 8.

Analysis of phage infection

Phage P1

E. coli MG1655 was used for routine phage maintenance. A P1vir phage lysate (6.3 ± 2.1 x10^9 PFU/ml) was prepared according to Ref. [44] and stored at 4°C for no longer than one

week. To examine the effect of CbAgo on cell viability, the kinetics of cell growth during infection with phage P1 was analyzed in strains with or without CbAgo. To obtain bacterial stocks, *E. coli* strains DE178 (MG1655 Z1::CbAgo-cat), DE179 (MG1655 Z1::dCbAgo-cat), DE182 (MG1655 Z1::noAgo-cat) were grown overnight in LB with chloramphenicol (25 μg/ml), diluted twice with 50% glycerol, aliquots were frozen in liquid nitrogen, stored at -80 °C and then used to obtain overnight bacterial cultures. The overnight cultures were inoculated into 25 ml of fresh LB medium supplemented with $CaCl_2$ (5 mM) and aTc (0.2 μg/ml) to OD_{600} = 0.05 and grown for 1 h at 30 °C with aeration. The phage P1 lysate was added to achieve the desired multiplicity of infection (MOI) of 0.1, 1, and 5, with a no-phage control, and the culturing was resumed. Cell density was monitored by measuring OD_{600} every 20 minutes on the Ultrospec 2100 pro UV Visible Spectrophotometer (GE Healthcare). Means and standard deviations of three biological replicates for each strain and MOI were calculated and plotted using a custom R script.

To determine phage titers during P1 infection, samples of bacterial cultures were taken at 160, 220 and 340 min after the start of infection. 30 μl of infected cultures was diluted 10-fold and vortexed with 100 μl of chloroform for 15 seconds with subsequent centrifugation. The supernatant was again treated with chloroform, the aqueous phase was transferred to a fresh tube and incubated with an open cap for 30 min. Serial dilutions were prepared in solution containing 10 mM $MgSO_4$, 5 mM $CaCl_2$ in a 96 well plate and plated in 15 μl aliquots on freshly prepared LB plates covered with top agar supplemented with 10 mM $MgSO_4$, 5 mM $CaCl_2$ and an MG1655 strain culture outgrown until OD_{600} = 0.2. The plates were air dried for 20 minutes, incubated overnight at 37 °C, and the phage plaques were counted.

To compare formation of plaques in strains with or without expression of CbAgo, 500 μl of freshly thawed cells of strains DE178, DE179 and DE182 were inoculated into 10 ml of liquid LB supplemented with chloramphenicol (25 μg/ml) and aTc (0.2 μg/ml) and grown at 20 °C at 220 rpm for 5.5 hours until OD_{600} = 0.2. The cells were collected by centrifugation and diluted in 3 ml of buffer containing 10 mM Tris-HCl, pH 7.4 and 10 mM $MgSO_4$, 100 μl aliquots were mixed with 100 μl of an appropriate phage P1 dilution in 10 mM $MgSO_4$, 5 mM $CaCl_2$ and the samples were incubated for 30 min at 30 °C. LB agar plates were prepared with 0.2 μg/ml aTc. Top agar (0.7%) was supplemented with 10 mM $MgSO_4$, 5

754 mM $CaCl_2$, 0.2 μg/ml aTc; 2.8 ml of top agar were mixed with 200 μl of the preincubated
755 mixture of bacteria with P1 and plated. The plates were incubated at 30 °C for 12-48 hours
756 to monitor morphology of phage plaques.
757 *Phage M13*
758 To prepare a stock of phage M13, an overnight culture of the *E. coli* NEB Turbo strain
759 (containing the F' factor required for M13 infection) was diluted 10-fold with 5 ml of LB
760 medium and a slice of agar with a plaque of phage M13 was added. The culture was
761 incubated for 4-5 hours at 37 °C with aeration. The cells were precipitated by
762 centrifugation (4000 g, 10 min) and the supernatant was filtered through a sterile 0.22 μm
763 syringe filter (Millex-GP, Merck, Millipore). The resulting phage stock was stored at +4 °C
764 for no more than one week. To determine the titer of phage M13, an overnight culture of *E.*
765 *coli* NEB Turbo was grown in liquid LB, diluted 100-fold with fresh LB and grown at 37 °C
766 with aeration until OD_{600}~0.5. A series of dilutions of phage M13 in liquid LB were
767 prepared, and 10 μl samples were mixed with 200 μl of cell cultures. The mixture was
768 incubated at room temperature for 1-5 minutes to adsorb phage particles on the cells,
769 mixed with 3 ml of molten (47 °C) 0.5% bacterial agar, and poured into a Petri dish with
770 solidified lower LB agar. The plates were incubated at 37 °C for 12-16 hours, and the
771 number of plaques was counted.
772 To analyze the effects of CbAgo on phage infection, cultures of *E. coli* NEB Turbo
773 Z1::CbAgo, Z1::dCbAgo, or a control strain without CbAgo (Supplementary Table 3) were
774 grown and aliquoted as described above for experiments on plasmid loss. Aliquots of
775 these cultures were added to LB with aTc (0.1 μg/ml) and grown overnight at 25 °C. 200 μl
776 of overnight cultures were inoculated into 20 ml of fresh LB+aTc, 10 μl of phage M13 stock
777 with the titer of 10^{12} was added, and grown at 25 °C with aeration. 5 ml aliquots were taken
778 after 2, 4, 6, and 8 hours and centrifuged (4000 g; 10 min). The supernatant was filtered
779 through 0.22 μm filters. Each sample was titrated in the standard way using the *E. coli*
780 NEB Turbo strain. For each strain, four biological replicates of the experiment were
781 conducted.
782
783 *Phage T7*

To obtain a phage T7 lysate, 0.25 ml of an overnight culture of *E. coli* MG1655 was inoculated into 5 ml of fresh LB medium and grown at 30 °C until OD_{600}=0.4. Then 100 μl of a phage T7 stock was added and the cells were grown until complete lysis. The lysate was treated twice with chloroform and stored at 4 °C. To determine the phage titer, serial dilutions of the phage lysate (100 μl) were added to an exponential culture of *E. coli* MG1655 (100 μl) at OD_{600}=0.6-0.8. The samples were incubated for 30 min at 37 °C, mixed with 3 ml of 0.8% top agar pre-heated at 55 °C and poured into a Petri dish with solidified lower LB agar. The number of plaques was counted after 12-16 at 30 °C.

To determine T7 titers during infection, overnight cultures of strains DE178 (MG1655 Z1::CbAgo-cat) and DE182 (MG1655 Z1::noAgo-cat), grown in LB containing aTc (0.1 μg/ml) at 25 °C, were inoculated into fresh LB with aTc and grown until OD_{600}=0.4. Different titers of phage T7 were added (1.5×10^7, 4×10^{10}, 1.5×10^{11} per 10 ml), the cultures were grown for 1-3 hours at 25 or 20 °C, treated with chloroform, and the phage titers were determined using *E. coli* MG1655 as described above.

Statistical analysis

For statistical analysis of phage infection, the data were $\log_{10}$-transformed, and then one-way ANOVA were carried out as an initial search for statistically significant differences between groups of data. P-values were calculated by applying Scheffe's test for multiple comparison of means. All statistical analysis was implemented in R (v. 3.6.1) using standard libraries.

Analysis of the distribution of pAgos, RecBCD/AddAB and CRISPR-Cas in prokaryotic genomes

The search for pAgo proteins in the recent NCBI protein database (downloaded in Dec 2019), their alignment, phylogenetic analysis and classification were performed as in Ref.[4]. Some strains have several NCBI accessible genomic assemblies, and in the order to remove such redundancy only the largest genomic assembly with the largest number of annotated proteins was considered for each strain. In total, we found 1,711 pAgos encoded in 2,883 genomes of 2,802 bacterial strains, including 833 pAgos that were not found previously[4]. For the construction of the PIWI-MID-based phylogenetic tree we

selected a subset of 399 pAgos of 456 strains with fully sequenced genomes ("Complete Genome" or "Chromosome" statuses).

The proteins belonging to the RecBCD and AddAB systems were searched for in the NCBI protein database (downloaded in Dec 2019) using *hmmsearch* (v. 3.2.1) and HMM profiles from the TIGR database and Ref.[45]: RecB (HMM profile RecB, TIGR00609); RecC (recC, TIGR01450); RecD (recD1, TIGR01447); AddA (addA_alphas, TIGR02784; addA_Gpos, TIGR02785; AddA_cremie and AddA_epsilon, Ref. [45]); AddB (addB_alphas, TIGR02786; addB_Gpos, TIGR02773; rexB_recomb; TIGR02774; AddB_cremie and AddB_epsilon, Ref. [45]). The protein was marked as the DSB repair protein if it had a match with any of these HMM profiles (E-value ≤ 1e-3). If the protein had matches with several HMM profiles, then the match with the best score was selected. Identified DSB repair proteins were attributed to NCBI genomic assemblies of prokaryotic strains (downloaded from NCBI FTP in Jan 2020), and then only 19,253 assemblies with "Complete Genome" or "Chromosome" statuses were further considered. The genome was considered as encoding RecBCD or AddAB systems if all proteins from the corresponding system were identified as encoded in this genome. RecBCD and AddAB were encoded in 7,419 (38.5%) and 8,141 (42.3%) prokaryotic genomic assemblies, correspondingly. Both RecBCD and AddAB were found in 56 (0.3%) genomes.

The search for Cas proteins and their classifications into the classes and subtypes were performed in the same set of full genomic assemblies using CRISPRCasFinder (v.4.2.18) [46]. Cas proteins from Class I and Class II systems were encoded in 6,413 (33.3%) and 1,647 (8.6%) assemblies, respectively; 450 (2.3%) genomes contained proteins from both classes.

Methods references

41. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6640-6645, (2000).
42. Bohn, C., Collier, J. & Bouloc, P. Dispensable PDZ domain of Escherichia coli YaeL essential protease. *Molecular microbiology* **52**, 427-435, (2004).
43. He, F. E. coli Genomic DNA Extraction *Bio-101*, e97, (2011).

44. Thomason, L. C., Costantino, N. & Court, D. L. E. coli genome manipulation by P1 transduction. *Current protocols in molecular biology* **Chapter 1**, Unit 1 17, (2007).
45. Bernheim, A., Bikard, D., Touchon, M. & Rocha, E. P. C. A matter of background: DNA repair pathways as a possible cause for the sparse distribution of CRISPR-Cas systems in bacteria. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **374**, 20180088, (2019).
46. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic acids research* **46**, W246-W251, (2018).
47. Khan, S. R., Gaines, J., Roop, R. M., 2nd & Farrand, S. K. Broad-host-range expression vectors with tightly regulated promoters and their use to examine the influence of TraR and TraM expression on Ti plasmid quorum sensing. *Applied and environmental microbiology* **74**, 5053-5062, (2008).
48. Guerry, P., van Embden, J. & Falkow, S. Molecular nature of two nonconjugative plasmids carrying drug resistance genes. *Journal of bacteriology* **117**, 619-630, (1974).
49. Chang, A. C. & Cohen, S. N. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *Journal of bacteriology* **134**, 1141-1156, (1978).
50. Bolivar, F. Construction and characterization of new cloning vehicles. III. Derivatives of plasmid pBR322 carrying unique Eco RI sites for selection of Eco RI generated recombinant DNA molecules. *Gene* **4**, 121-136, (1978).
51. Studier, F. W. & Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of molecular biology* **189**, 113-130, (1986).

Data and Code Availability

All data generated during this study are included in the published article and the Extended Data and are available from GEO database with the accession number GSE148596. The code used for data analysis is available at the GitHub repository https://github.com/AntKuzmenko/CbAgo_DNAi.git

Acknowledgements

We thank Dr. Martin A. White for strains with engineered I-SceI sites, Dr. Gerry Smith for RecBCD antibodies, Sergey Lysenkov for the help with statistical analysis and Anna Olina for the help with the phage experiments. D.Y. would like to thank Matvey Kolesnik for helpful discussions. The project was in part supported by the Ministry of Science and Higher Education of the Russian Federation (14.W03.31.0007), Russian Science Foundation (19-14-00359), Russian Foundation for Basic Research (18-29-07086). A.A.A. is supported by an HHMI Faculty Scholar Award. D.L. is supported by grant MR/M019160/1 from the MRC (UK).

Author information

Affiliations

Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia
Anton Kuzmenko, Anastasiya Oguienko, Daria Esyunina, Denis Yudin, Mayya Petrova, Alina Kudinova, Olga Maslova, Sergei Ryazansky, Alexei Aravin, Andrey Kulbachinskiy

Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
Anton Kuzmenko, Maria Ninova, Alexei Aravin

Institute of Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3FF, UK
David Leach

Contributions

A.Kuz., A.A.A. and A.Kul. conceptualized the study. A.Kuz., D.Y. and D.E. constructed strains. A.Kuz., A.O., D.Y. and D.E. prepared smDNA libraries. A.Kuz., A.O. and D.E. prepared genomic DNA libraries. A.Kuz., A.O. and D.Y. analyzed sequencing data, M.N. and S.R. helped with data analysis. S.R. performed phylogenetic analysis. D.L.

conceptualized experiments with engineered DSBs. A.Kuz., A.Kud., O.M., M.P., A.O. and D.E. performed experiments on plasmid elimination and phage infection. All authors interpreted the results. A.Kuz. and A.O. prepared the figures. A.Kul. and A.A.A. wrote the manuscript with contribution from other authors.

Competing interest declaration

The authors declare no competing interests.

Supplementary Information is available for this paper.

Corresponding authors

Correspondence and requests for materials should be addressed to Anton Kuzmenko, Alexei A. Aravin and Andrey Kulbachinskiy.

Extended Data Fig. 1. Small DNAs associated with CbAgo.

(a) Analysis of small nucleic acids isolated from wild-type CbAgo and dCbAgo. The samples were treated with alkaline phosphatase, ^{32}P-labelled with polynucleotide kinase and treated with DNase I (D), RNase A (R) or left without further treatment (-). CbAgo is associated with small DNAs, as confirmed by their sensitivity to DNase treatment and resistance to RNase. The DNA marker (M) lengths are indicated. For gel source data, see Supplementary Figure 1.

(b) Length distribution of smDNAs associated with CbAgo in the wild-type, *recBrecD*, *recC* and *recA* strains. For the *recC* strain, there is a small increase in the smDNA length, suggesting that their processing might be different in this strain.

(c) and (d) Analysis of nucleotide biases for chromosomal (wild-type CbAgo and dCbAgo), plasmid (pNonChi) and phage M13 smDNAs associated with CbAgo. (c) Nucleotide frequencies at different guide positions. (d) AT/GC-content along the guide length and in surrounding genomic sequences. Guide positions starting from the 5'-end are indicated below the plots. For genomic DNA, the AT-bias around the first position is seen for both active CbAgo and dCbAgo. The AT-bias in the downstream region (positions 14-18) is seen for active CbAgo but not for dCbAgo. For each replicon, the average GC-content of smDNAs corresponds to the GC-content of this replicon (shown in percent in each panel), indicating that the efficiency of smDNA processing does not strongly depend on the GC-content.

(e) Model of processing of smDNAs by CbAgo from double-stranded DNA precursors. Binding of the guide 5'-end in the MID-pocket of CbAgo may be facilitated by melting of the DNA duplex in the upstream guide region (left scheme). Guide DNA loading is completed after CbAgo-dependent cleavage of the complementary DNA strand and its dissociation, depending on the AT-content of the downstream guide-target duplex (right scheme).

Extended Data Fig. 2. Whole-genome mapping of smDNAs associated with CbAgo in strains with various genetic backgrounds.

(a-h) For each strain, the distribution of smDNAs along the chromosome is shown in RPKM (the number of smDNAs reads per kilobase per million reads in the smDNA library). Left panels show total smDNA counts, right panels show strand distribution of smDNAs for each strain (plus DNA strand, green; minus DNA strand, red). Positions of the *araC*, *lacI* and *ter* sites are shown above the plots. SmDNA coverage is shown in RPKM. The identities of the strains and plasmids, with plasmid or chromosomal localizations of the CbAgo gene, are indicated (see Supplementary Tables 2 and 3). (a) Wild-type BL21(DE3) with plasmid-encoded CbAgo (pBAD containing the *araC* gene). (b) The same in BL21(DE3) with knockout of Tus. (c) MG1655Z1 with genomic CbAgo. (d) The same with pET28b containing *lacI*. (e) Plasmid-encoded catalytically dead dCbAgo in BL21(DE3). (f) Knockout of RecB/RecD in BL21(DE3) with plasmid-encoded CbAgo. (g) Knockout of RecC in BL21(DE3) with plasmid-encoded CbAgo. (h) Knockout of RecA in BL21(DE3) with plasmid-encoded CbAgo. The observed enrichment of smDNAs around the *ori* region

in the *recC* and *recA* strains may possibly reflect the higher DNA content and/or a higher likelihood of DSB formation in this region in these strains.

(i) Targeting of specific genomic regions depends on the catalytic activity of CbAgo. The ratio of smDNAs between wild-type CbAgo and dCbAgo (obtained for BL21(DE3) containing corresponding pBAD_CbAgo plasmids) is shown in the logarithmic scale. Normalized densities of smDNA reads (RPKM) were calculated for each CbAgo variant and plotted as a WT/dCb ratio. The regions with the ratio of >1 correspond to the sites of active smDNA processing by CbAgo. CbAgo targets the *araC* locus, *ter* region and multicopy sequences: rDNA operons (indicated with arrows above the plot) and IS elements. Positions of IS1 (29 copies) and IS3 (12 copies) in the BL21(DE3) genome are shown with dotted lines below the plot.

Extended Data Fig. 3. **Asymmetry in smDNA distribution at specific genomic loci.**

(a) Zoomed-in peaks of smDNAs around the *araC* and *lacI* genes in strains containing plasmids with corresponding genes. (b) Examples of smDNA distributions around rRNA operons, *rrsD* and *rrsC*, in wild-type *E. coli* and strains with knockouts of *recBrecD* and *recC*. The reads from the plus and minus genomic strands are shown in green and red, respectively. Positions of Chi sites in surrounding genomic regions are indicated (forward for the plus strand and reverse for the minus strand); the closest Chi sites in the corresponding strands (forward for the plus strand, reverse for the minus strand) are shown with dotted lines.

(c) Metaplot of the number of smDNAs around Chi sites in each genomic strand (red, plus-strand smDNAs for plus-strand Chi sites; green, minus-strand smDNAs for minus-strand Chi sites) in the 2-3 Mb genomic region. Position around Chi is shown in kilobases.

(d) Strand-specific asymmetry in smDNA distribution for various strains (ratio of RPKM values for the plus and minus genomic strands). A similar bias is observed for the wild-type and *recBrecD, recC* and *recA* mutant strains expressing CbAgo but not in wild-type cells expressing catalytically inactive dCbAgo.

Extended Data Fig. 4. Growth kinetics of *E. coli* strains depending on the expression of CbAgo.

Growth kinetics of *E. coli* BL21(DE3) and its mutant derivatives with or without CbAgo (containing pBAD_CbAgo or empty pBAD plasmids) at 30 °C in the rich LB (a) and minimal M9 (b) media. Overnight cultures of cells were inoculated into fresh LB to OD_{600} of 0.01 in the presence of the inductor (0.05% L-arabinose) and cell density was measured at 15 min intervals in a microplate reader.

Extended Data Fig. 5. Whole genome analysis of DNA content in the wild-type and *tus*$^-$ *E. coli* strains depending on the expression of CbAgo. The experiment was performed with wild-type (a,b) or *tus*- (c,d) BL21(DE3) containing or lacking the pBAD_CbAgo plasmid. The cells were harvested at the exponential phase (OD_{600}=0.5) (a,c) or stationary phase (OD_{600}=6) (b,d), followed by isolation of total DNA and sequencing. For each condition, genomic DNA coverage is shown for strains without and

with CbAgo, and the ratio for the +CbAgo and -CbAgo strains is shown in a separate panel (black). The enlarged *ter* region and the *araC* locus are shown separately. Genomic DNA coverage is shown in RPKM. At the stationary phase, a peak in genomic DNA coverage was detected in the strains containing CbAgo, which exactly corresponded to the DE3 prophage in BL21(DE3). This may indicate formation of DSBs in this region, possibly as a result of partial prophage excision, leading to DNA repair and replication.

Extended Data Fig. 6. Targeting of engineered DSBs by CbAgo.

(a) *Top,* smDNA abundance in the chromosomal area spanning the engineered DSBs (palindrome or I-SceI-dependent; I-SceI^{mut}, the mutated cleavage site) and *ter* sites, for the wild-type CbAgo or dCbAgo. In each strain, the numbers of smDNAs mapping to the region of DSB are shown in percent of total smDNAs. The presence of the DSB shifts the ratio between the *terA* and *terC* peaks in favor of *terA*, likely as a result of impediment of the clockwise replisome, moving toward *terC*, by the DSB formation. *Bottom,* strand-specific distribution of smDNAs around engineered DSBs for strains with dCbAgo (palindrome and I-SceI DSBs) or with wild-type CbAgo and the I-SceI^{mut} site. The reads from the plus and minus DNA strands are shown in green and red, respectively. Most smDNAs are produced from the 3'-strand at each end of the DSB, and the boundaries of the smDNA peaks are defined by Chi sites. (b) Genomic DNA coverage in the same region in palindrome-containing strains depending on the expression of active CbAgo or dCbAgo. (c) The ratio of genomic DNA profiles for palindrome-containing strains with wild-type CbAgo and dCbAgo relative to the strain without CbAgo. Wild-type CbAgo but not dCbAgo triggers DNA loss around the DSB with overreplication of genomic DNA at the site of termination. (d) Genomic DNA coverage at DSBs formed by the I-SceI meganuclease in *E. coli* strains with induced I-SceI but without expression of CbAgo (left) and with expression of both I-SceI and CbAgo (right). (e) The ratio between genomic DNA profiles for the strains with and without expression of CbAgo. Genomic DNA coverage is shown in RPKM.

Extended Data Fig. 7. Targeting of plasmid and phage DNA by CbAgo.

(a) SmDNA coverage of plasmids (pNonChi, left; pBAD_CbAgo, right) in strains with plasmid-encoded CbAgo. The moving average of smDNA coverage in a 200 nt window is shown for the plus and minus DNA strands (green and red, respectively). (b) SmDNA coverage of a pET28 plasmid in a strain with chromosomal CbAgo. (c) Distribution of smDNAs along the M13 genome. SmDNAs were isolated from CbAgo expressed in *E. coli* NEB Turbo strain during infection with M13. (d) Coverage of plasmid DNA in whole-genome sequencing in the wild-type and *tus*$^{-}$ strains, depending on the expression of CbAgo. The values represent the moving average of genomic DNA coverage in a 200 nt window (in RPM).

(e-g) Targeting of the F' plasmid by CbAgo. Total smDNA coverage (e), coverage of the plus and minus DNA strands (f) and the plus to minus strand ratio (g) are shown along the F' sequence. Positions of the three copies of IS3 element, the origin of replication (*oriS*), the core part of the F factor, and the chromosomal insertion ('chr') are indicated. For

strand-specific smDNA distribution, positions of the nearest Chi-sites in the corresponding strands are shown. The vast majority of reads map to the F episome core sequence lacking Chi sites, and the numbers of smDNAs drop significantly upon encountering the first Chi site. The distribution is also asymmetric relative to the origin of replication. Similarly to the chromosome (Extended data Fig. 3d), the lagging DNA strand is targeted with a higher efficiency, suggesting a connection to replication.

Extended Data Fig. 8. Loss of plasmids after various number of passages in *E. coli* strains with or without CbAgo.

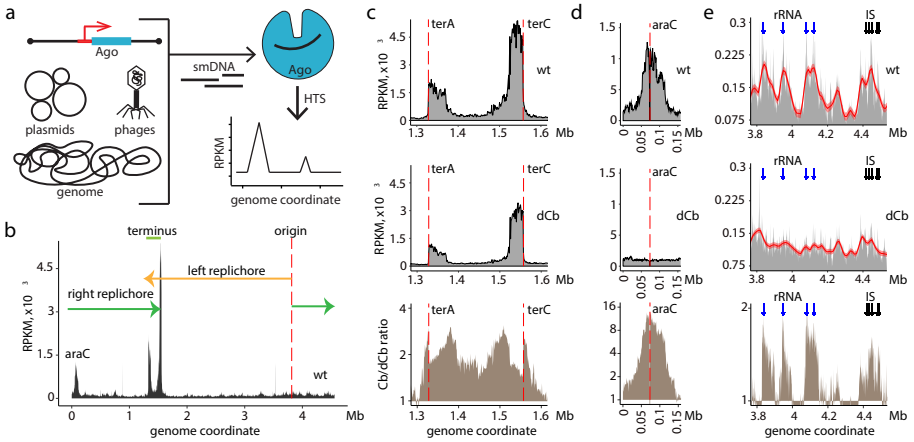
Cells expressing genome-encoded CbAgo (Cb), its catalytically dead mutant (dCb) or without Ago ('without') were transformed with one of the six different plasmids from different incompatibility groups. Percentage of plasmid-free cells were measured after indicated number of passages (means and standard deviations from 2-4 biological replicates). CbAgo, but not dCbAgo facilitates plasmid elimination regardless of the plasmid type.

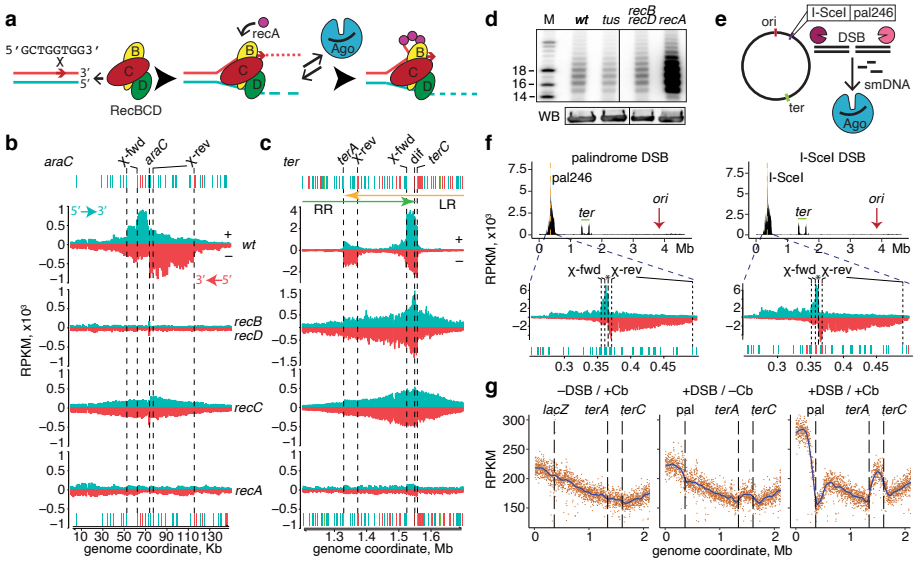
Extended Data Fig. 9. Effects of CbAgo and dCbAgo on P1 infection.

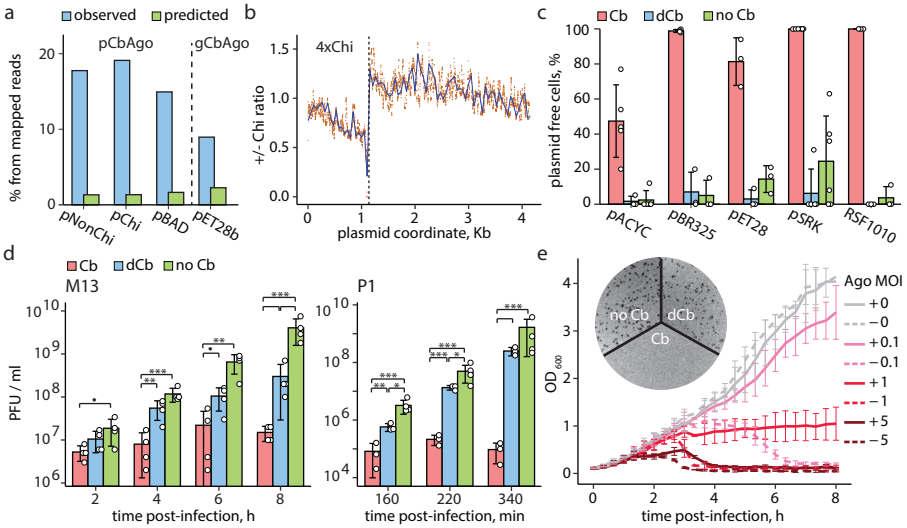
(a) Bacterial culture growth during P1 infection with different MOI in strain with dCbAgo. Means and standard deviations from three independent experiments. (b) Titers of P1 at MOI 1 and 5 at different times post-infection in strains without CbAgo or with expression CbAgo or dCbAgo. PFU, plaque forming units. Means and standard deviations from three-four independent measurements (*$p<0.05$, **$p<0.01$, ***$p<0.001$).

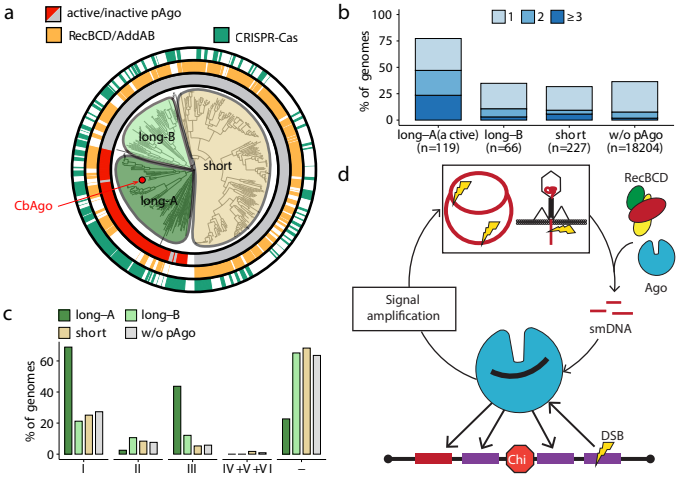
Extended Data Fig. 10. Co-occurrence of pAgo proteins, DSB repair systems and CRISPR-Cas in prokaryotic genomes.

(a) Circular phylogenetic tree of pAgos from prokaryotic strains with fully assembled genomes based on the multiple alignment of the MID-PIWI domains. Three major phylogenetic groups of pAgos are indicated (see Ref.[4]): 'long-A' pAgos usually contain all characteristic domains of the Ago family (N, PAZ, MID and PIWI) and have a predicted nuclease site; 'long-B' also contain all domains but are inactive; 'short' pAgos contain only MID and PIWI domains and are inactive. The pAgo proteins were annotated as follows, from the inner to the outer circles: the superkingdom to which the corresponding pAgo belongs; the type of the PIWI domain, depending on the presence of the catalytic tetrad DEDX; the type of the DSB repair system encoded in the corresponding genome; the class of CRISPR-Cas system; the type and subtype of CRISPR-Cas system. CbAgo, SeAgo (*Synechococcus elongatus* Ago) and TtAgo (*Thermus thermophilus* Ago) are highlighted in red. The scale bar represents the evolutionary rate calculated under the JTT+CAT evolutionary model. (b) The distribution of various subtypes of Type I and Type III CRISPR-Cas systems in the fully assembled genomes encoding pAgos. The number of genomes for each pAgos group is indicated.


a
Ago
smDNA
Ago
HTS
plasmids
phages
genome
RPKM
genome coordinate
b
terminus
origin
left replichore
right replichore
araC
wt
RPKM, x10 3
genome coordinate
c
terA
terC
wt
dCb
Cb/dCb ratio
RPKM, x10 3
genome coordinate
d
araC
wt
dCb
genome coordinate
e
rRNA
IS
wt
dCb
genome coordinate
Mb


a
5′GCTGGTGG3′
X
3′
5′
RecBCD
recA
Ago
B
C
D
d
M
wt
tus
recB
recD
recA
18
16
14
WB
e
ori
I-SceI
pal246
DSB
smDNA
ter
Ago
b
araC
X-fwd
araC
X-rev
5′→3′
3′←5′
wt
recB
recD
recC
recA
RPKM, x10³
genome coordinate, Kb
c
ter
terA
X-rev
X-fwd
dif
terC
RR
LR
genome coordinate, Mb
f
palindrome DSB
pal246
ter
ori
RPKM, x10³
X-fwd
X-rev
Mb
I-SceI DSB
I-SceI
g
–DSB / +Cb
+DSB / –Cb
+DSB / +Cb
lacZ
terA
terC
pal
RPKM
genome coordinate, Mb


a
observed
predicted
pCbAgo
gCbAgo
% from mapped reads
pNonChi
pChi
pBAD
pET28b
b
4xChi
+/- Chi ratio
plasmid coordinate, Kb
c
Cb
dCb
no Cb
plasmid free cells, %
pACYC
pBR325
pET28
pSRK
RSF1010
d
M13
P1
PFU / ml
time post-infection, h
time post-infection, min
e
no Cb
dCb
Cb
Ago MOI
+0
−0
+0.1
−0.1
+1
−1
+5
−5
OD 600
time post-infection, h


a
active/inactive pAgo
RecBCD/AddAB
CRISPR-Cas
long-B
short
long-A
CbAgo
b
1
2
≥3
% of genomes
100
75
50
25
0
long–A(a ctive) (n=119)
long–B (n=66)
short (n=227)
w/o pAgo (n=18204)
c
long–A
long–B
short
w/o pAgo
% of genomes
60
40
20
0
I
II
III
IV +V +VI
–
d
RecBCD
Ago
Signal amplification
smDNA
DSB
Chi