

# Natural C-independent expression of restriction endonuclease in a C protein-associated restriction-modification system

Monika Rezulak, Izabela Borsuk and Iwona Mruk\*

Department of Microbiology, University of Gdansk, Wita Stwosza 59, 80-308 Gdansk, Poland

Received February 23, 2015; Revised November 12, 2015; Accepted November 13, 2015

## ABSTRACT

**Restriction–modification (R-M) systems are highly prevalent among bacteria and archaea, and appear to play crucial roles in modulating horizontal gene transfer and protection against phage. There is much to learn about these diverse enzymes systems, especially their regulation. Type II R-M systems specify two independent enzymes: a restriction endonuclease (REase) and protective DNA methyltransferase (MTase). Their activities need to be finely balanced *in vivo*. Some R-M systems rely on specialized transcription factors called C (controller) proteins. These proteins play a vital role in the temporal regulation of R-M gene expression, and function to indirectly modulate the horizontal transfer of their genes across the species. We report novel regulation of a C-responsive R-M system that involves a C protein of a poorly-studied structural class - C.Csp231I. Here, the C and REase genes share a bicistronic transcript, and some of the transcriptional auto-control features seen in other C-regulated R-M systems are conserved. However, separate tandem promoters drive most transcription of the REase gene, a distinctive property not seen in other tested C-linked R-M systems. Further, C protein only partially controls REase expression, yet plays a role in system stability and propagation. Consequently, high REase activity was observed after deletion of the entire C gene, and cells bearing the  $\Delta$ C R-M system were outcompeted in mixed culture assays by those with the WT R-M system. Overall, our data reveal unexpected regulatory variation among R-M systems.**

## INTRODUCTION

Many species of bacteria and archaea possess restriction-modification (R-M) systems (1,2) that, among other roles, serve to resist bacteriophage predation and to modulate

gene flow (3,4). R-M systems are classified into four main types, with Type II numerically dominating Types I, III and IV (5). A type II R-M system is typically composed of two independent enzymes: a restriction endonuclease (REase) that cleaves DNA at a specific sequence, and a modification methyltransferase (MTase) that acts on the same sequence to protect it from cleavage by the cognate REase. This gives a simple mechanism for discriminating between self- and non-self DNA (6,7) in that the REase can degrade DNA entering the cell, while resident host DNA remains uncleaved due to methylation by the MTase. Type II R-M system activity may be harmful to its host, when expression of the REase and MTase is not finely balanced. Toxic REase action may cause double-strand breaks in a genome, and it can be lethal for the host if unrepaired (8,9). It resembles the functioning of many toxin-antitoxin units and other genetic addiction elements, which demand a counterbalance to toxicity to avoid the post-segregational killing of the host (10–12). Since R-M systems are highly abundant and mobile within Prokaryotes, (5,13) some mechanisms must exist to provide the coordinated, temporal control of REase expression. It is especially crucial during the transfer of an R-M system to a new host cell wherein, the genome is unprotected by methylation. A sufficient strategy employed is to delay REase expression to allow the MTase first to complete genome methylation (14). The molecular basis of these processes is still enigmatic, and it is understood only to some extent for Type II R-M systems. It seems likely that they require fine-tuned transcriptional feedback circuits to keep the REase/MTase activities in balance.

Three control mechanisms for Type II R-M system expression have been outlined: MTase, antisense RNAs, and C proteins, but none of these mechanisms is understood in great detail. In the first case, a MTase represses its own transcription via cognate operator binding within the promoter (15–17). Additionally, in some R-M systems the MTase recognition sequence for modification is located within its own gene's promoter, and the methylated promoter sequence represses MTase gene expression. Eventually the low level of enzyme leads to loss of methylation after replication and promoter activity is unblocked (18,19). Such

\*To whom correspondence should be addressed. Tel: +48 58 5236071; Fax: +48 58 5236074; Email: shamrock127@hotmail.com

feedback loops operate to prevent MTase overexpression, but do not explain the control of toxic REase expression. For the Ecl18kI R-M system, the control of expression is exerted at the level of transcription kinetics when MTase autorepression is accompanied by an additional promoter competition mechanism to ensure the MTase/REase expression is balanced (20).

The second mechanism of coordinated expression of R-M system seems to apply to bicistronic systems where the REase gene precedes the MTase, as in EcoRI and Eco29kI (21–24). In these cases antisense RNAs produced from oppositely-oriented promoters negatively regulate REase and MTase expression (23,24). Weakening or knocking out an antisense promoter results in enhancement of REase expression to the point of toxic enzyme accumulation and cell death (23). In contrast, overexpression of antisense RNA *in trans* alleviates REase toxicity after loss of the R-M system (23).

The third important mode of R-M system expression control relies on a specialized protein transcription factors called C proteins. These controller proteins are present in more than 300 R-M systems, and were first discovered in the PvuII and BamHI R-M systems (25,26). Typically, a C gene precedes an REase gene and sometimes partially overlaps it (26–35). The upstream location of C facilitates efficient and precise C-dependent transcriptional control over the toxic REase gene (35–38). In one tested exception to this paradigm, Kpn2I, the C gene precedes and controls the MTase gene, but has no effect on REase gene expression (39). In one of the best-studied C-dependent R-M systems, PvuII, the C and REase genes share the same bicistronic transcript controlled by both a weak C-independent promoter and a stronger C-dependent promoter (29,36,40). If C protein is inactivated or absent then REase expression is very low, and REase activity is undetectable. If the C gene is supplied *in trans* then REase levels return to the wild-type values (29,34). The C/REase expression level is a result of a gene-copy-dependent feedback loop that activates transcription at low C protein concentrations and represses when levels become high (36,37,41). To exert control C protein binds to a specific palindromic DNA sequence (C-box) that is embedded in its own promoter region. In the Esp1396I R-M system, there is a C-box upstream of the C gene, and another upstream of the MTase gene; each has distinct C binding affinities (38). In another case, the C protein is translationally fused to the REase, and it controls the expression of the fused gene (42). More generally, C proteins are grouped into three informal prototype families (C.PvuII, C.EcoRV and C.EcoO109I) based on the conservation of the C-box nucleotide sequence. The C.EcoO109I family has been studied the least (36), and more systematic detection of inverted repeats/palindromes upstream of C genes has identified several additional motifs (43,44).

This study focuses on a new regulatory class of C protein-associated R-M systems with the C protein (C.Csp231I) being a member of the understudied C.EcoO109I family. A crystal structure for C.Csp231I is available (45,46), but *in vivo* regulatory studies have not been performed for related R-M systems except in the C-box class prototype EcoO109I (33). We find that the C protein role in Csp231I is distinct from its role in EcoO109I, and we report a novel regulatory

mode for a C protein-associated R-M system with separate promoters for the C and REase genes. We further demonstrate high REase activity, regardless of C gene presence, which is unlike the case with other characterized C proteins. We do discuss the possible role that C.Csp231I plays in regulating its cognate R-M system, but in general we find that much remains to be learned even about this subset of C-controlled R-M systems.

## MATERIALS AND METHODS

### Strains, plasmids and oligonucleotides

The source of studied R-M system was *Citrobacter* sp. RFL231, and was kindly supplied by Dr. A. Janulaitis, MBI Fermentas, Lithuania. Despite the fact the *E. coli* and *Citrobacter* are both members of the *Enterobacteriaceae*, expression of the Csp231I R-M system in *E. coli* cells seemed to be toxic. Accordingly, to clone the WT Csp231I R-M system the competent cells were prepared from *E. coli* MM294 strain, which expresses the MTase gene from pEcoVIII M to ensure the protection of host genome. M.Csp231I and M.EcoVIII both recognize the same nucleotide sequence (47,48). Other *E. coli* K-12 strains used in this study are described below. MC1061 [*araD139*  $\Delta$ (*ara, leu*)7697,  $\Delta$ *lacX74*, *galU*, *galK*, *hsdR*, *strA*] was used in all *lacZ* reporter assays (49). *E. coli* DH5 $\alpha$  and MM294 were used for all other purposes including cloning steps. *E. coli* Rosetta was used for C protein overproduction and purification (50). The plasmids used are listed in Supplementary Table S1. They were also deposited in the Collection of Plasmids and Microorganisms, University of Gdansk, Poland. The oligonucleotides used are shown in Supplementary Table S2 of the Supplementary Data.

### Effect of C gene delivered *in trans*

The gene for C.Csp231I or its variant was PCR-amplified and cloned downstream of the arabinose-inducible P<sub>araBAD</sub> promoter in vector pBAD24 or pBAD33 (51) yielding series of wildtype (pBAD-CWT) and variant (pBAD-arq and pBAD-sqe) plasmids (Supplementary Table S1). Arabinose induction experiments were performed in M9-minimal medium with 0.2% glycerol as the carbon source (52). Briefly, single colonies were used to inoculate overnight cultures in M9 media supplemented with appropriate antibiotics. These cultures were diluted 1:50 into antibiotic free medium, and grown with shaking to an OD<sub>600</sub> of 0.2–0.3. The cells were then gently pelleted, resuspended, and divided among flasks containing M9-minimal media with varied concentrations of L-arabinose. After about 3 hours of subsequent cultivation, ONPG assay was performed and Miller or modified Miller units were calculated as previously described (23,36,53). A series of transcriptional *lacZ* fusions were generated in the pRS415 vector (54), and translational *lacZ* in-frame fusions were created using pLex3B (55). Details regarding plasmid cloning and features are outlined in the Supplementary Data (Supplementary Table S1).

### RNA isolation, RT-PCR and determination of transcription start points

*E. coli* carrying p18 plasmid with WT Csp231I R-M system were grown to exponential phase and pelleted. Total cellular RNA was isolated using the Total RNA Kit (A&A Biotechnology, Poland). The transcription start points of the genes encoding the Csp231I R-M system were determined by the primer extension method. The 5' ends of appropriate primers were labeled with 5 pmol of [ $\gamma$ -<sup>32</sup>P]ATP (primer EXMET for mapping P<sub>M</sub>; primer EXRES for mapping P<sub>R</sub> and C16 for mapping P<sub>C</sub>). Twenty- $\mu$ l primer extension reactions containing 10  $\mu$ g RNA, 0.6 pmol of labeled primer, buffer (50 mM Tris-HCl pH 8.3, 50 mM KCl, 4 mM MgCl<sub>2</sub>, 10 mM DDT), 1 mM of each dNTPs and 10 u RiboLock RNase Inhibitor were denatured at 95°C for 3 minutes, and then incubated at 50°C for 1 hour. Next, 200u RevertAid H Minus Reverse Transcriptase (Fermentas) was added and samples were incubated at 42°C for 30 min. Finally, 4  $\mu$ l of loading dye (95% formamide, 0.05% bromophenol blue, 0.05% xylene cyanol) was added, samples were denatured at 75°C and then loaded onto 6% acrylamide:bis (19:1)–7 M urea in 1 $\times$  TBE gel (52). Sequencing reactions were also performed on DNA templates using the DNA Cycle Sequencing Kit (Jena Bioscience), and the appropriate radiolabelled primer. These samples were loaded on the acrylamide gels with the primer extension reactions.

For reverse transcription PCR (RT-PCR), 5  $\mu$ g of total RNA was DNase I treated in solution at 37°C for 1 h using the RNase-free DNase I (Eurx, Poland). After a 20 min inactivation at 65°C, the cDNA synthesis was performed using RevertAid H Minus Reverse Transcriptase (Fermentas) kit with random hexamers according to the manufacturer's protocol. The resultant cDNA was then used as template for PCR with indicated primer pairs. The resultant PCR products were separated on 2% TBE agarose gels.

### Electrophoretic mobility shift assays (EMSA)

DNA specific substrate was double-stranded PCR-amplified (primers: C20–C24; p18 plasmid as a template, Supplementary Tables S1 and S2) fragments that were 576 bp in length and included the entire P<sub>C</sub> promoter/operator region. The non-specific DNA substrate (515 bp) containing the P<sub>M</sub> promoter region was amplified with C2-C30 primers. Reactions containing 100 nM DNA and the indicated protein concentrations were prepared in binding buffer [50 mM Tris-HCl (pH 8.0), 1 mM DTT, 10 mM MgCl<sub>2</sub>] to a final volume of 20  $\mu$ l, and incubated for 20 min at 22°C. Samples were electrophoresed on 6% native polyacrylamide gels in 0.5 $\times$  TBE buffer at 22°C. The location of dsDNA in the gels was determined by ethidium bromide staining.

### Overproduction and purification of the C protein and its variants

The coding sequence of C gene was PCR-amplified from *Citrobacter* sp. RFL231 genomic DNA using primers CNco and CRev. NcoI and EcoRI treated PCR fragments were cloned into pET28(+) linearized with the same restriction

enzymes to generate pET-CWT, which produces a C protein with a C-terminal His5-tag. The other C gene variants were generated using Quick-Change mutagenesis on the pET-CWT template. Resultant plasmids were: pET-Csqe (primers sqe1 and sqe2) producing a C protein variant with the following triple amino acid residue substitution S16A; Q17A; E18A, and pET-Carq (primers arq1 and arq2) producing the A33G; R34E; Q37A C protein variant. The vectors were sequenced confirmed and are called C-SQEmut and C-ARQmut respectively throughout the text.

For protein purification, the host *E. coli* Rosetta strain was used to overexpress all three His5-tagged C protein variants (50). Cells were grown in 100 ml LB broth supplemented with appropriate antibiotics at 37°C. At an OD<sub>600</sub> of 0.3 C protein production was induced by adding IPTG to 0.5 mM. After a 3 hour incubation, cells were pelleted and stored at –70°C until used. Frozen cells were thawed in C buffer (50 mM NaH<sub>2</sub>PO<sub>4</sub>, 300 mM NaCl, 10 mM imidazole, 5% glycerol, 1 mM PMSF and 5 mM  $\beta$ -ME pH 8.0) and sonicated (40  $\times$  10 s). The lysates were cleared by centrifugation and applied to a column packed with TALON cobalt resin (Clontech), and washed with buffer C. The bound proteins were eluted with 150 mM imidazole in C buffer. Protein-containing fractions were pooled, and dialysis was performed over night in buffer containing 50 mM MgCl<sub>2</sub>, 5% glycerol, 50 mM Tris-HCl pH 8.0 to remove excess of imidazole. Finally, purified protein samples were concentrated in 50% glycerol and stored at –20°C. The protein concentration was determined by densitometry in Tricine-SDS-PAGE (56) using a lysozyme as a quantitative standard.

### Western blot analysis

Samples of cultures containing similar number of cells were centrifuged, supernatants were removed, and the cell pellets stored at –80°C. Pellets were resuspended in 1 $\times$  SDS Laemmli buffer (52), and lysed by heating at 98°C for 10 min. Proteins were resolved by Tricine-SDS PAGE (56) and then electroblotted to PVDF membranes. C.Csp231I protein bands were detected by chemiluminescence using the ECL-plus Western Blotting Detection System (GE Health Sciences) with 1:2000 dilution of rabbit anti-C.Csp231I polyclonal serum prepared according to standard protocols (57), and a 1:30 000 dilution of horseradish peroxidase-conjugated goat anti-rabbit IgG. Protein bands were visualized either by autoradiography or by using 5-bromo-4-chloro-3-indolylphosphate (BCIP) as the alkaline phosphatase substrate and nitroblue tetrazolium (NBT) as the color development reagent. The prestained MW markers used were PageRuler (Fermentas).

### Efficiency of transformation assay

Efficiency of transformation (EOT) is defined in this study as the relative number of transformants obtained from a given preparation of competent cells using a non-saturating amount of plasmid DNA. EOT is calculated from the ratio of transformants with a test plasmid relative to those with the control vector. This term is equivalent to the term 'relative transformation efficiency'. In this particular case

due to toxicity of the investigated R-M system, *E. coli* cells MM294 carrying plasmid with the protective MTase gene [pEcoVIII] were used as a competent cells. The standard CaCl<sub>2</sub>-heat shock method of transformation was used (52).

### Relative restriction activity assay

The restriction activity of *E. coli* cells carrying Csp231I R-M system and its variants was measured by determining plating efficiency of bacteriophage  $\lambda$ vir. The efficiency of plaquing (plaque-formation) (EOP) of  $\lambda$ vir was calculated as the ratio of plaques formed on *E. coli* MM294 [pEcoVIII] containing plasmid pBR322 (restriction negative) to those formed on the same strain containing a plasmid with the Csp231I R-M system or its variants. Relative restriction (= 1) refers to the WT R-M system.

### Fitness assay by mixed culture competition experiment

Five ml of M9-glucose medium without any antibiotic was inoculated with two comparably sized colonies of *E. coli* MG1655 obtained from a fresh transformation grown on LB-agar plates with appropriate selective antibiotic. One colony carried the p18amp plasmid bearing the WT R-M system, and was selective on ampicillin. The second colony, selected on tetracycline, contained its variant - the p30tet plasmid wherein the C gene and its C-box sequence were deleted. The *bla* gene of p30tet was disrupted to change the plasmid selection marker to tetracycline resistance. In parallel control cultures, plasmids with restriction-negative and modification-positive variants (p17amp versus p31tet) or empty vectors (pBRamp versus pBRtet) were used. Time zero of the competition experiment marks the point at which 1:1 mixed cultures of competing cells were inoculated. Every 15–18 hrs of incubation at 37°C with shaking, the co-cultures were diluted 10<sup>6</sup> into fresh minimal media without antibiotics. A sample of each mixed competition culture was immediately taken, appropriately diluted and spread quantitatively either onto LB-agar containing the appropriate selective antibiotic or on antibiotic free LB agar. The colonies were counted and the ratio of colony-forming units (CFU) of the two competing cell populations was calculated using  $T = (\text{CFU}_{\text{amp}}/\text{CFU}_{\text{tet}})$ ; their generation number was also determined. Data were normalized using the results from the vector control ( $T = (\text{CFU}_{\text{amp}}/\text{CFU}_{\text{tet}})$ ). The relative competitive fitness ( $W$ ) was then calculated as  $W = \log[T/V]$  for each tested generation time-point (58).

## RESULTS

As reported here, the Csp231I R-M system was cloned from chromosomal DNA of *Citrobacter* sp. RFL231, by selecting for the MTase. We used a suicide plasmid that carried a functional REase gene (no MTase gene), from which cells would be protected if they had a MTase with corresponding specificity (59). The recognition sequence (5'-AAGCTT-3') is the same as the prototype HindIII R-M system (60) as well as several other isospecific systems studied in our laboratory: EcoVIII from *E. coli* E1585–68 (47,48,61); LlaCI from *Lactococcus lactis* subsp. *cremoris* W15 (62,63) and BstZ1II from *Bacillus stearothermophilus* 14P (59).

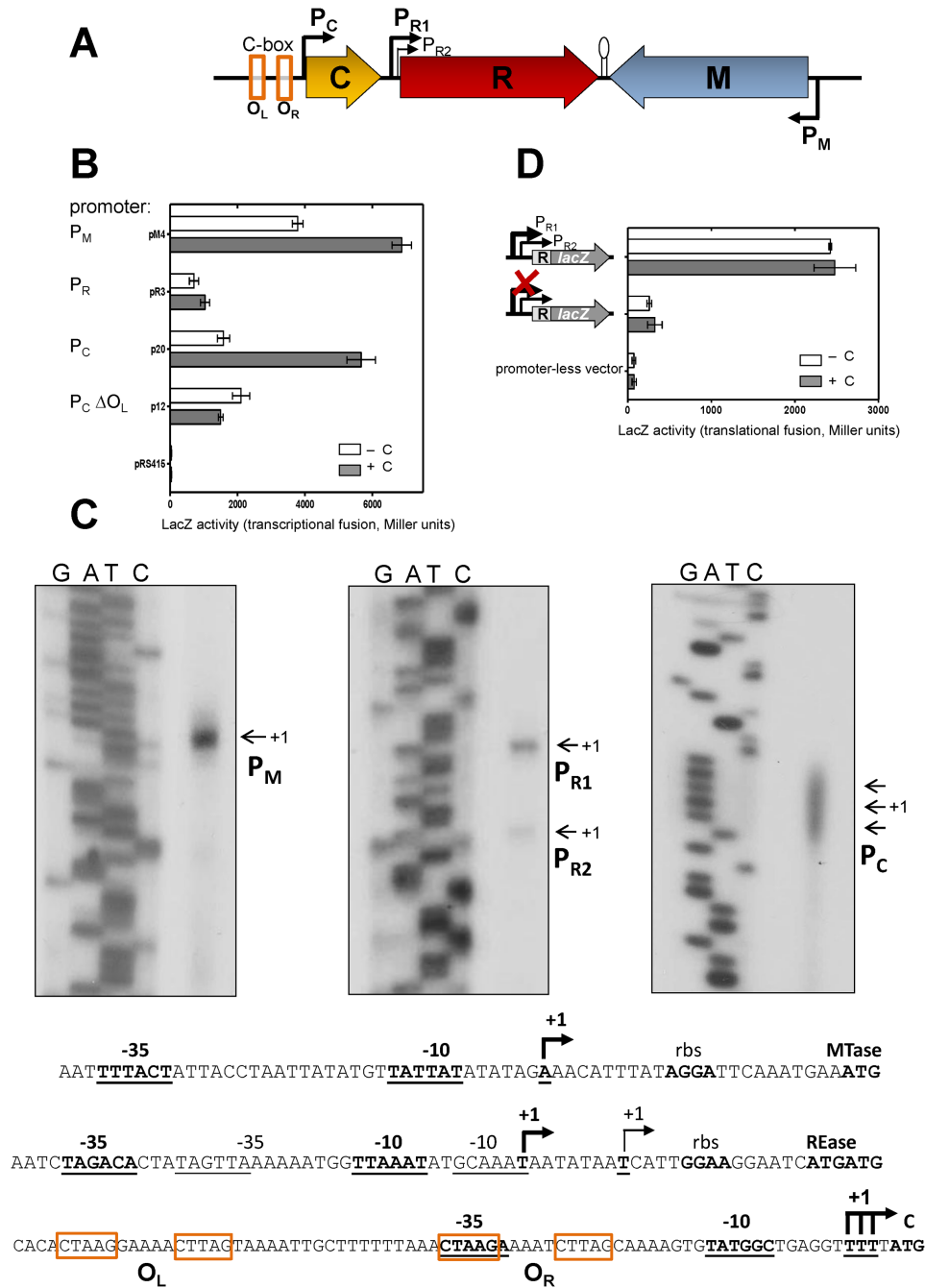
The genetic organization of Csp231I R-M system is shown in Figure 1A. In addition to the convergently oriented REase (*csp231IR*) and MTase (*csp231IM*) genes, the regulatory C gene (*csp231IC*) is present. The *csp231IC* gene is upstream of *csp231IR* in a typical colinear location. The genes do not overlap, unlike other C and REase genes (31,64), which reduces the possibility of translational coupling. Moreover, the sequence analysis indicated a potential Rho-independent transcription terminator in the 152-nt intergenic region separating *csp231IR* and *csp231IM* genes ([www.softberry.com](http://www.softberry.com)).

### Each gene of the Csp231I R-M system has its own promoter

To determine the localization of promoters, we cloned the upstream regions of the three Csp231I R-M system genes into the pRS415 vector upstream of a promoterless *lacZ* reporter gene. LacZ enzyme activity was measured via ONPG hydrolysis assays. Promoter activity was detected for all three DNA fragments in: pM4 (P<sub>M</sub> of MTase gene), p20 (P<sub>C</sub> of C gene) and pR3 (P<sub>R</sub> of REase gene) (Figure 1B; white bars). The latter result was unexpected as REase genes in C protein-associated R-M systems typically rely on a shared promoter upstream of C gene (29,36,40).

Next, we tested whether the detected promoter activities change when C.Csp231I is delivered *in trans* (Figure 1B; dark gray bars). The C gene was cloned into a compatible plasmid under a P<sub>araBAD</sub> promoter to generate pBAD-CWT. The resulting LacZ activities revealed a positive regulatory effect of C protein on both the P<sub>M</sub> and P<sub>C</sub> promoters, but surprisingly not on P<sub>R</sub> (Figure 1B; black bars). The P<sub>C</sub> and P<sub>M</sub> promoter activity increased about 3.5 and 2.0-fold, respectively. The regulatory effect of a C protein is related to its binding to inverted repeats (C-box; motif 8; (43)) located usually within the promoter region. We found two such repeats and deleted the left part (O<sub>L</sub>) generating plasmid p12. The promoter activity this variant remained unchanged regardless of the C protein presence or absence. These results indicate there is at least one promoter that can be stimulated for higher activity (directly or indirectly) by C protein. In the putative promoter for the MTase gene (P<sub>M</sub>), no inverted repeats have been detected. A similar observation of the C protein effect on MTase expression of unclear origin was reported previously (33).

To localize the promoters, we identified transcription start sites of Csp231I genes by primer extension, using total RNA prepared from *E. coli* cells carrying the complete, functional Csp231I R-M system. For the MTase gene, a single primer extension product was produced (Figure 1C). Probable appropriately positioned –10 (TATTAT) and –35 (TTTACT) sequences were identified based on comparison to *E. coli* consensus sequences (TATAAT and TTGACA with 17nt spacer, respectively (65)). For the REase gene, two primer extension products were found with one being much more intense, and so assumed to be the major promoter (P<sub>R1</sub>; Figure 1C). The identified –10 (TTAAAT) and –35 elements (TAGACA) of P<sub>R1</sub> each showed only a single nucleotide difference from the consensus sequences. The minor P<sub>R2</sub> promoter revealed –10 and –35 elements as GCAAAT and TAGTTA. To test the REase promoter identification the TTAAAT –10 box of the major P<sub>R1</sub> pro-



**Figure 1.** Csp231I R-M system and its transcription initiations. (A) Genetic organization of the Csp231I R-M system comprising of its regulator (C gene and its promoter P<sub>C</sub>), REase (R gene and its two promoters: major P<sub>R1</sub> and minor P<sub>R2</sub>) and MTase (M gene and its P<sub>M</sub>) (not to scale). The identified promoters are designated by arrows. The operator for P<sub>C</sub> promoter, C-box, consists of two inverted repeats CTAAG-n<sub>5</sub>-CTTAG, left and right, marked as O<sub>L</sub> and O<sub>R</sub> respectively. A presence of Rho-independent transcription terminator in the 152-nt intergenic region separating R and M genes is predicted and depicted here as hairpin. (B) Identified promoters activity was measured as a transcriptional fusion of appropriate DNA fragments with indicated promoter to the reporter *lacZ* gene. In case of P<sub>C</sub>, deletion of left operator (ΔO<sub>L</sub>) also has been tested. The transcription activity in context of C protein absence (vector control, white bars) or C protein presence (pBAD-CWT, dark bars) is presented in Miller units. The results are the averages (±SD) of at least three independent experiments. (C) Mapping the transcription start sites for the promoters with confirmed activity. For each reaction, total RNA from *E. coli* harboring the p18 plasmid with entire, functional R-M system was used as template for primer extension method using radioactively labeled primers and performed as indicated in Materials and Methods. The primer extension products (marked as +1) were resolved on a denaturing 6% polyacrylamide gel along with the nucleotide sequencing reactions (G, A, T, C) performed with the same labeled primer and appropriate DNA template. At the bottom, sequence of relevant DNA containing the indicated promoters is also shown. The -10 and -35 promoter motifs are underlined and the start codons (ATG) and ribosome binding sites are in bold. In case of P<sub>C</sub> promoter, the sequence of two inverted repeats CTAAG-n<sub>5</sub>-CTTAG is boxed. For REase promoters, the start of transcription for major promoter P<sub>R1</sub> is indicated by thicker arrow and bold -10 and -35 boxes, in contrast to the minor P<sub>R2</sub> promoter marked by thin arrow. (D) Promoter activity for REase gene was tested by ONPG hydrolysis in plasmid constructs of pLEX3B, where reporter *lacZ* gene was fused in-frame to REase gene (pLEX-PR<sub>1</sub> WT). The major P<sub>R1</sub> promoter (thicker bent arrow) was knocked-out by mutation of -10 box of TTAAGT→CCCGGG (pLEX-PR<sub>1</sub> mut). The transcription activity in context of C gene presence or absence was measured as in panel B.

motor was changed to CCCGGG (pLEX-P<sub>R1</sub> WT versus pLEX-P<sub>R1</sub> mut) (Figure 1D), and this drastically decreased transcription activity regardless of C gene presence *in trans*. However, the remaining activity was still slightly above the promoterless level (pLex3B) probably indicating weak activity from the second P<sub>R2</sub> promoter (Figure 1D).

Mapping the start of mRNA for the C gene showed a single spot of multiple products within a stretch of five 'T's located on the DNA template just upstream of the translational start (Figure 1C). This indicates that the transcript is leaderless (the start codon is either preceded by only a few nucleotides, as in this case, or it starts directly with a 5-terminal AUG) much like many homologous regulatory C genes (27,35,40). The multiple products from the primer extension most likely result from reiterative transcription that may occur during transcription initiation, and can impact gene expression in some cases (66,67). Nevertheless, we found these sequences suboptimal to the consensus sequence: CTAAGA (−35 box) and TATGGC (−10 box). Since we considered a second transcript for the C gene to be possible, we carefully mapped a start site on RNA isolated from cells carrying the p20 reporter gene in the presence and absence of C (as in Figure 1B), and each time the same products within a single cluster were obtained (Figure 1C and data not shown). We also confirmed the absence of other transcripts for the C gene by reverse transcription PCR and it indicated that no mRNA is produced from the sequence upstream of −10 position of P<sub>C</sub> promoter (Supplementary Figure S1).

### C protein specifically binds its C-box sequence *in vitro*

We next tested if the observed stimulation of transcription in presence of the C gene is due to C protein binding the inverted repeats (C-boxes) located upstream of the C gene. To date, most of the tested C proteins (C.PvuII-like) recognize two operators comprising nearly palindromic sequences of GACT-tat-AGTC, separated by a highly conserved central spacer with GT conserved at the center (36,68). C.Csp231I, which represents a new class of C proteins, was grouped based on the unusual sequence of the two inverted repeats, CTAAG-n<sub>5</sub>-CTTAG separated by an extended AT-rich 18-nt spacer, that showed no obvious similarity to the binding site of its homolog C.EcoO109I (33,45). Moreover, the C.Csp231I crystal structure revealed additional two helices at the C-terminus (H6 and H7; Figure 2A) that may play a role in dimer interface stabilization (45). Even with these differences the positional alignment of two large groups of C proteins (C.PvuII-like and C.Csp231I-like) still revealed highly similar short regions in the amino-acid sequence (Figure 2A, gray bars). With the exception of the regions linked to transcription activation and DNA recognition, no tests were performed to assign amino acid conservation to C protein features. The structural comparison of C.PvuII and λcI repressor (40) shows the common amino-acid region with Glu (E) residue as a vital contact to σ<sup>70</sup> of RNA polymerase to achieve transcription activation (69). We also found a conserved region (G<sub>14</sub>LSQE<sub>18</sub>) in the C.Csp231I amino-acid sequence (Figure 2A), and generated a C.Csp231I protein variant in which S<sub>16</sub>Q<sub>17</sub>E<sub>18</sub> of helix H2 were replaced; S16A; Q17A; E18A (designated C-SQEmut).

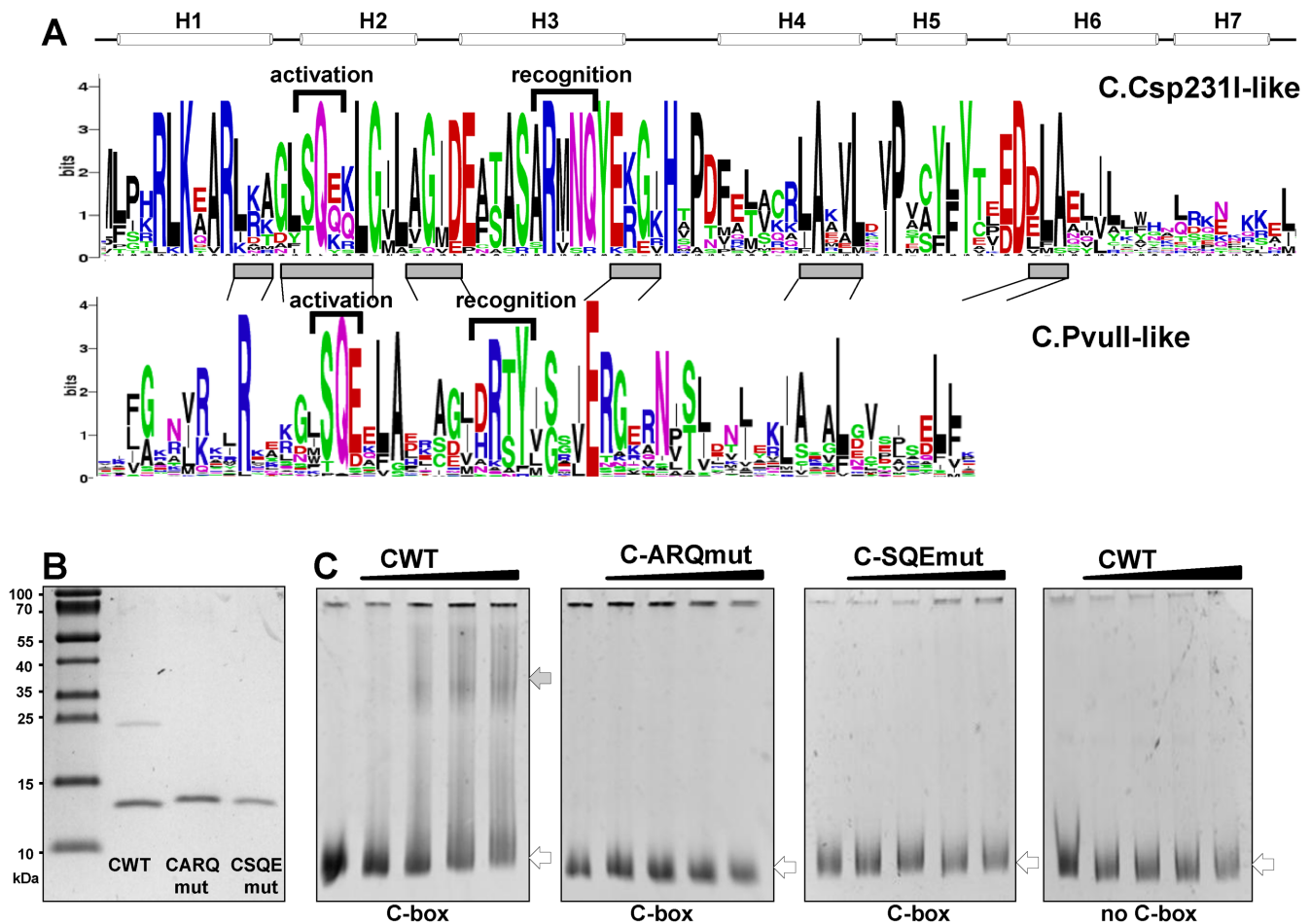
We also constructed a C protein variant (A33G; R34E; Q37A) that is predicted to impair DNA binding (45) by inhibiting the engagement of H3 in DNA recognition (Figure 2A). We designated this C variant as C-ARQmut. Both variants, as well as the WT protein C.Csp231I, were expressed and purified to homogeneity (>95%) as C-terminal His-tag fusion proteins (Figure 2B). The predicted molecular mass of 12 046 Da for C.Csp231I-His matched the value estimated from SDS-PAGE. Interestingly, the C-WT fused to His-tag, unlike its changed variants, showed two bands on the SDS-PAGE gel that correspond to the monomer (more intense) and dimer even under the denaturing conditions used during the protein separation (Figure 2B). The identity of these two forms of C.Csp231I was confirmed by western blot (Supplementary Figure S2). This observation supports the presence of a strong dimerization interface for C.Csp231I that was previously predicted from the crystal structure (45).

Next, we tested whether or not the C.Csp231I-His and its variants bind to the DNA fragment containing the operator bearing both of the inverted repeats of C-box sequence *in vitro* (boxed sequences of O<sub>L</sub> and O<sub>R</sub> in Figure 1C, bottom). EMSA reactions were performed with the same amount of a 576-bp target DNA (100 nM) and increasing concentrations of the different C proteins (0–2000 nM) (Figure 2C). DNA of comparable size and with non-specific sequence (no C-box) was used as a control in testing the specificity of the DNA-protein interaction. The specific shift in DNA-protein migration was observed only for C.Csp231IWT, but not for its variants: C-ARQmut or C-SQEmut (Figure 2C). The retarded complex was not distinct and some unbound DNA still remained. Different target DNAs (biotin-labeled 32-nt oligonucleotides) were also tried and they produced similar results (Supplementary Figure S3). Overall, these data indicate ARQ and SQE substitutions in C protein heavily disturb the interaction between the regulator and its operator within the inverted repeats.

In addition, we have noted a sequence between the MTase promoter hexamers (Figure 1C), which weakly resembles the single repeat of C-box recognized by C.Csp231I (Supplementary Figure S4). We performed EMSA reaction and did not obtain any retarded complex. C protein binding was not observed at least in tested conditions (Supplementary Figure S4).

### C.Csp231I positively and negatively regulates its own P<sub>C</sub> promoter

In general, C proteins are dedicated transcription factors that activate and repress their own transcription. We examined if C.Csp231I also acts accordingly using a previously tested *in vivo* titration of C protein to verify (36). Our experimental system is based on two compatible plasmids, one of which produces the C protein in a controlled fashion (pBAD-CWT or its variants) while the second plasmid carries a reporter gene (*lacZ*) fused to the C.Csp231I operator/promoter sequence (p20), which is the target for C protein action. C expression is controlled by the *araBAD* promoter, which is repressed by glucose and induced by arabinose over a wide range (51). Thus, C protein level can vary from undetectably low to high levels as shown by west-



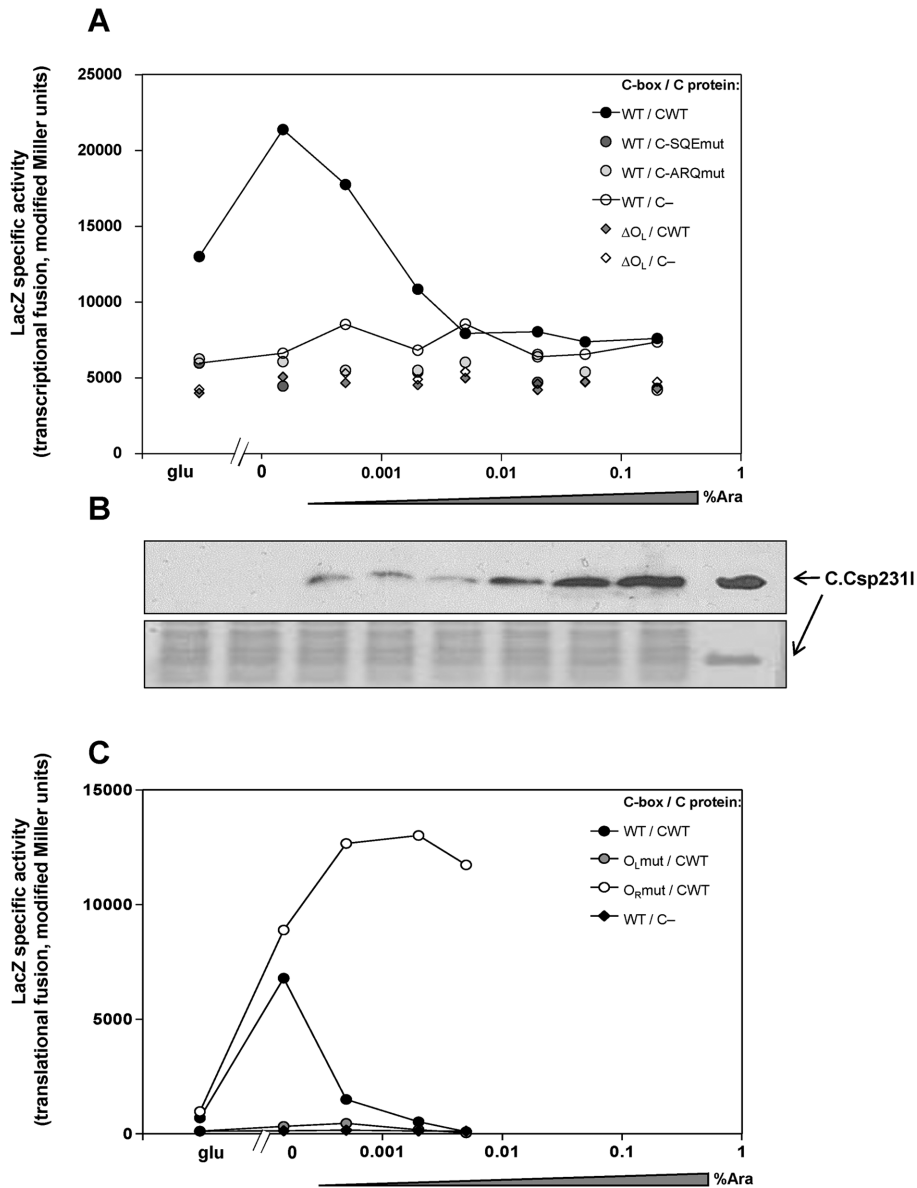
**Figure 2.** C.Csp231I protein and its action *in vitro*. (A) LOGO of C proteins sequence for: C.Csp231I family (21 sequences, upper part); C.PvuII family (65 sequences, bottom part). Helical structure, designated as seven white tubes, is derived from C.Csp231I crystal analysis (45). The regions of proteins responsible for activation of transcription and DNA recognition are inferred based on structural analysis of C.AhdI, C.Csp231I or mutational analysis of other C.PvuII-like members. Grey boxes represent well conserved residues with the highest similarity regions between the two C protein families: C.PvuII and C.Csp231I. Logo was generated from software at <http://weblogo.berkeley.edu>. (B) Preparations of C-terminal His-tag fusion of WT C protein (3  $\mu$ g) and its variants: C-ARQmut (2.8  $\mu$ g) and C-SQEmut (2  $\mu$ g), resolved on a 10% acrylamide Tricine SDS gel and Coomassie Blue stained. Overproduction and purification were carried out as described in Materials and Methods section. (C) C.Csp231I and its variants binding to the  $P_C$  promoter/operator region (C-box) containing the two inverted repeats CTAAG-n<sub>5</sub>-CTTAG. A 576-bp target DNA fragment was prepared by PCR amplification, as well as its control with no C-box sequence (515-bp) containing a DNA fragment of comparable size but lacking the C-box (Supplementary Table S2). Each binding reaction was carried out with the same amount of DNA (100 nM) and increasing concentration of proteins: 0, 200, 500, 1000 and 2000 nM. Reactions were processed further as outlined in Materials and Methods and finally resolved on 6% native polyacrylamide gels. DNA was visualized by ethidium bromide staining. Open and filled arrows denote positions of unbound DNA and protein–DNA complexes, respectively. The comparable data were obtained with modified protocol (Supplementary Figure S3).

ern blotting (Figure 3B). The host bacteria were *E. coli* MC1061, which is deficient in *lac* and *ara*, but able to transport arabinose. Experiments were performed in minimal media and the effect of C protein dependent transcription from  $P_C$ -*lacZ* was assessed by ONPG hydrolysis.

The results reveal that only WT C protein can activate the transcription from WT  $P_C$  promoter/operator yielding about a 4-fold increase compared to activity without C protein (Figure 3A). Interestingly, the peak of activity was achieved when no arabinose was added to the cells, and the C protein level was very minimal (undetectable by western blot) likely indicating expression leakage of  $P_{araBAD}$ . Further increase in the C protein concentration were associated with progressive LacZ reduction, which reached a level seen without C protein at 0.002% arabinose. This demonstrates

that, like several other C proteins, C.Csp231I acts as both an activator and a repressor of transcription. As a control, we used several combinations of C protein and C-box variants (WT and with  $O_L$  deleted; Figure 3A). These *in vivo* test of the C variants C-ARQmut or C-SQEmut confirmed our *in vitro* data (Figure 2C), as none of them could induce transcription activity above the no-C protein level, which stayed similar across the full range of arabinose concentrations.

To determine which operator ( $O_L$  and  $O_R$ ) upstream of the  $P_C$  promoter containing the two inverted repeats CTAAG-n<sub>5</sub>-CTTAG is associated with activation or repression (see Figure 1C, bottom sequence) we made separate mutations exchanging CTTAG for GTATC in each operator in pLex15, which has the C gene and its regulatory region fused to *lacZ*. The experiments were performed as



**Figure 3.** C.Csp23II is a transcriptional positive and negative autoregulator. Cells were grown in minimal media with 0.2% glucose (glu) or arabinose (ara) at indicated concentration from 0–0.2%. Cells contained two plasmids: one with WT promoter or its variant fused to reporter gene, second, compatible plasmid delivered WT C gene under inducible  $P_{araBAD}$  promoter, its C variant or C-absent vector. (A) Transcriptional profiles were measured as LacZ specific activity. Circles represent values for WT C gene promoter/ operator including two inverted repeats ( $O_L$  and  $O_R$ ) fused to *lacZ* reporter gene in p20 plasmid. Diamonds represent values for C gene promoter with  $O_L$  operator deleted and fused to *lacZ* reporter gene in p12 plasmid. As a source of C inducible expression following plasmids were used: pBAD-CWT; pBAD-sq; pBAD-arq or pBAD33 vector as a C-negative control. For plot clarity, C.Csp23II effect on its own transcription is represented by connected points: black circles in C presence and open circles in C absence. (B) Crude extracts from cells showing WT C-dependent transcription profile indicated by black circle in panel A, were resolved by 10% acrylamide Tricine-SDS PAGE and analyzed by western blotting using the rabbit anti-C.Csp23II polyclonal antibodies to detect C protein level (visualized by chemiluminescence). Bottom part serves as a protein loading control stained by coomassie brilliant blue. (C) Effect of altered C-box on C-dependent transcription profile. Mutation within  $O_L$  or  $O_R$  were introduced (CTTAG→GTATC; pLex15- $O_R$ mut or pLex15- $O_L$ mut) and measured as translational LacZ activity in WT C protein presence (circles; pBAD-CWT) or C absence (diamond, pBAD33). Transcriptional (panel A) and translational (panel C) activity was measured as LacZ specific activity determined by linear regression of the slopes for the lines generated by plotting LacZ activity versus optical density of the culture (modified Miller units) (53). Error of each point was measured with R-squared value not less than 0.94. Note, LacZ activity is from a transcriptional fusion in panel A, and from a translational fusion in panel C.



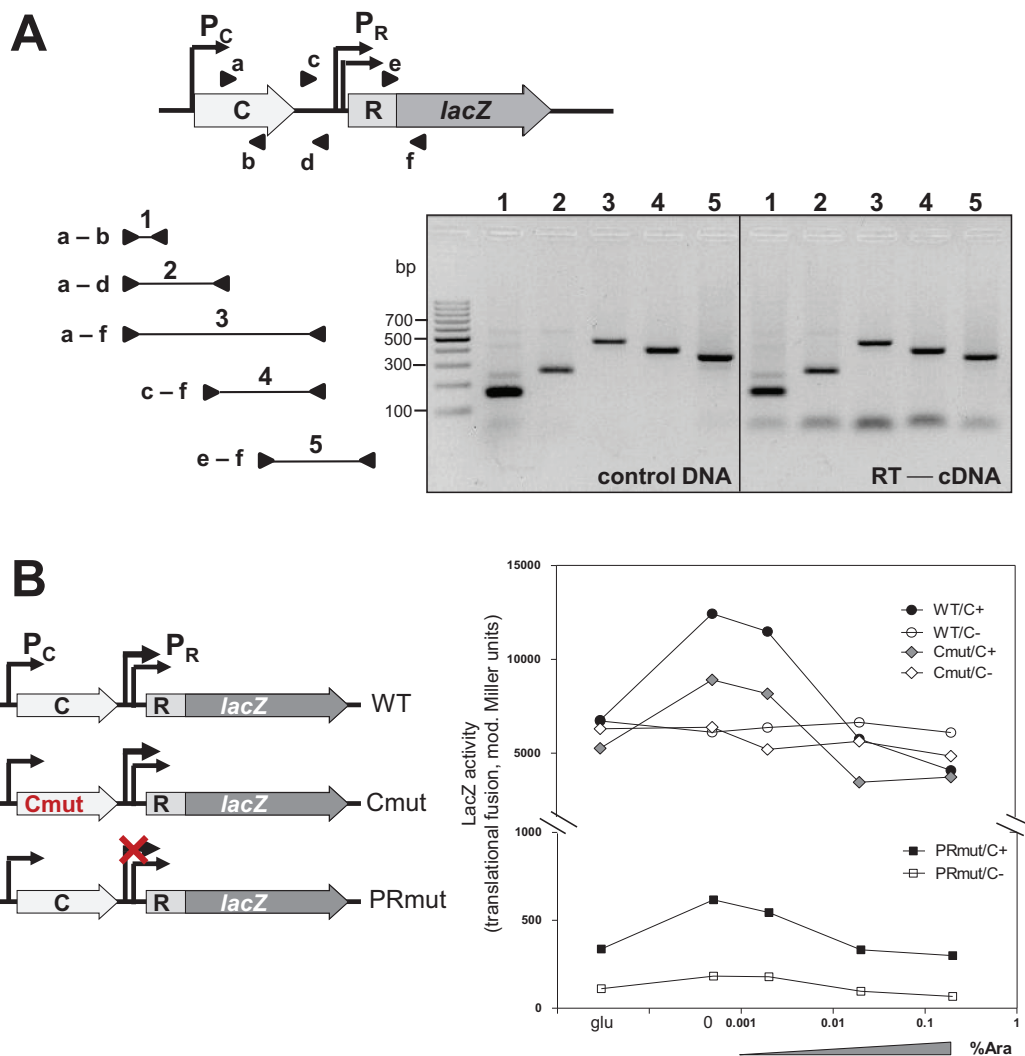
above with a second plasmid carrying the WT C gene under control of  $P_{araBAD}$  or with the empty vector as a control. The results confirmed our prediction of a common regulatory pattern for C proteins. Mutation within  $O_L$  abolished the activation arm in the transcription profile yielding *lacZ* activity similar to that in the C-negative control (Figure 3C). In contrast, mutation within  $O_R$  leads to comparable activation to WT C-box, but results in loss of transcriptional repression in the presence of C protein (Figure 3C).

### C and REase genes share a bicistronic transcript

For all previously tested systems with C proteins (e.g. PvuII, EcoRV, AhdI, Esp1396I), a bicistronic C-REase mRNA initiating from the C gene promoter has been found. While our data for Csp231I indicate that the REase gene has its own

promoter, it did not rule out possible readthrough transcription of the REase gene from the C promoter. To test whether this occurred we used reverse transcription PCR (RT-PCR) with primers encompassing the C/R region. PCR products were generated for each primer set located within the C or REase genes clearly indicating that a bicistronic transcript is made (Figure 4A).

Next, we wanted to test if  $P_C$  and  $P_{R1,2}$  contribute to REase expression equally or not. We used similar approach to the reporter assay discussed above (Figure 3), but in this case the reporter *lacZ* gene was inserted downstream to measure the REase expression level (Figure 4B). We quantified the REase transcriptional profile for a C protein gradient for WT (pLex-WT), for a variant defective in C protein binding to C-box (pLex-Cmut; as ARQ→GEA), and for another variant with an inactive  $P_{R1}$  (pLex-PRmut).



**Figure 4.** Transcripts initiated from  $P_C$  are bicistronic. (A) Schematic diagram of the C and R::*lacZ* genes in tested plasmid (not to scale). Triangles depict the primers (a–f) used in PCR reaction, where cDNA reversely transcribed from *E. coli* harboring pLexWT total RNA (right panel) or control pLexWT DNA (left panel) was used. The expected PCR products were: 1 (150bp); 2 (230bp); 3 (500bp); 4 (399 bp) and 5 (300 bp). (B) Transcription activity measured downstream of  $P_R$  promoter. The reporter *lacZ* gene was fused in-frame to REase gene generating a series of plasmids, which important features are indicated on the diagrams (red in left panel; not to scale). Briefly: Cmut indicates a substitution: A33G; R34E; Q37A in C protein;  $P_{R1}$  promoter was knocked-out as in Figure 1D. The transcription activity in context of a second plasmid carrying no C gene (pLex3B vector, white bars) or C gene (pBAD-CWT, dark bars) is presented in modified Miller units, as in Figure 3. Error of each point was measured with R-squared value not less than 0.95.

$P_{R1}$  knock out showed a drastic effect on REase expression (WT/C+ versus PRmut/C+) making it clear that the major effect on REase gene expression is via  $P_{R1}$  and not the C gene (Figure 4B). The mutation abolishing C binding to C-box resulted in slightly lower REase expression than WT (WT/C+ versus Cmut/C+) (Figure 4B). For the  $P_{R1}$  knock out we still see the effect of  $P_C$  driving the bicistronic mRNA as revealed by the peak of REase expression in presence of C protein at same concentration as for WT. We also noted that the delivery of WT C gene *in trans* for the variant disabled in C-box binding (Cmut/C+) did not fully restore the WT REase expression. Instead the REase expression was significantly increased at lower C concentration (Cmut/C+ versus Cmut/C-) (Figure 4B). Overall, this leads to the conclusion that C protein does control the REase expression, but the effect is only partial in magnitude making it unlike the other C-linked RM systems studied to date.

### C protein presence apparently is not essential for restriction endonuclease expression

To further understand the role of C protein and other genetic elements in the modulation of Csp231I R-M system, the genetic variants already tested by LacZ translational fusion assays were implemented in the context of the entire R-M system (Figure 5, schematic maps). First, their biological effect was determined by plaque formation assay with use of  $\lambda$ vir bacteriophage, which carries six recognition sites for R.Csp231I. Changes to amino acid residues within the C protein that prevent binding to its C-box sequence (S<sub>16</sub>Q<sub>17</sub>E<sub>18</sub>→AAA, p23 and A<sub>33</sub>R<sub>34</sub>Q<sub>37</sub>→GEA, p28; Figure 5A) resulted in ~8- and 11-fold decreases in restriction activity in comparison to the WT R-M system (p18) (Figure 5A). Similarly, a variant C-box sequence with deleted O<sub>L</sub> (p19) also showed about 4-fold less restriction activity. The knockout of the major REase promoter (p32) had the most pronounced effect in reducing restriction activity at about 100 times less than WT, which is also in accord with our reporter assay (Figure 4B). Surprisingly, however, a deletion of the entire region upstream of  $P_R$  that included the entire C gene and its linked operator (p30) resulted in the highest phage restriction, that is, about three times that of the WT Csp231I R-M system (Figure 5AB). This *in vivo* observation apparently indicates that C protein is dispensable for the restriction activity of the Csp231I R-M system.

### The C gene-absent R-M system exhibits impaired establishment in new host cells

Typical C proteins act as ‘timing regulators’ to delay the appearance of toxic REase activity in new unprotected hosts during R-M system transfer (36). We next tested if the C protein is required for Csp231I R-M system establishment in a new host. The plasmids carrying the WT R-M system or its variants were introduced into cells, and the relative efficiency of transformation (EOT) was determined (Figure 5C). For some R-M systems high restriction activity contributes to poor establishment of the restriction plasmid due to the lethal effect on the acceptor cells. This is true of the p18 plasmid carrying the WT Csp231I R-M system (Figure 5C). Among its variants, only a REase-negative mutant

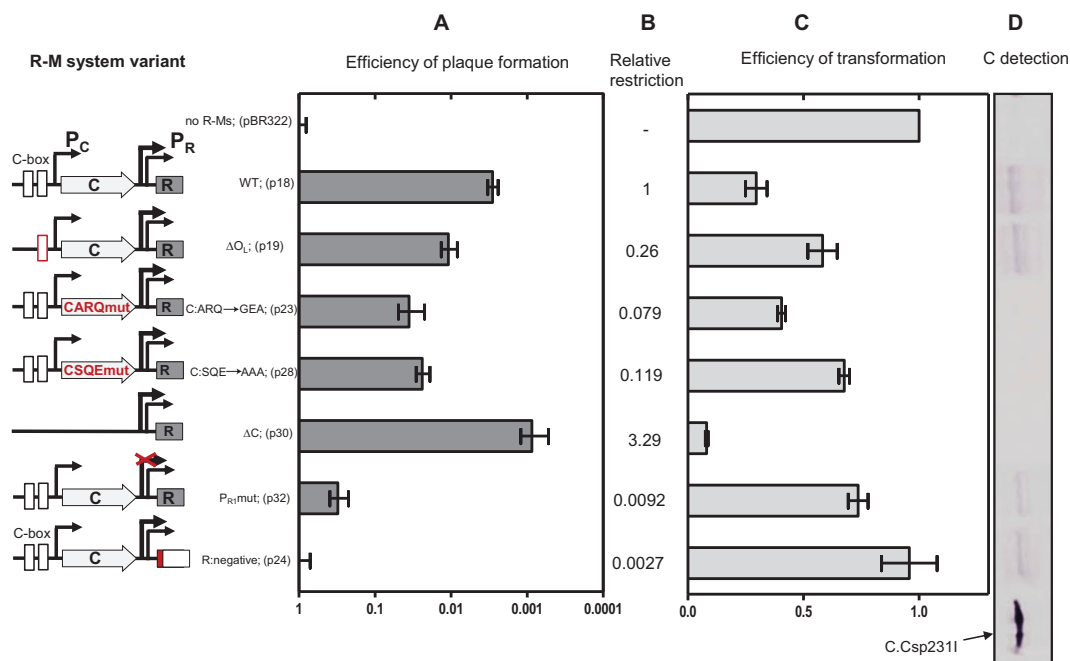
(p24) showed a high transformation efficiency comparable to that of the control plasmid (pBR322). The wild-type and other variants had varied establishment defects that usually correlated with their cellular restriction activity level. The R-M system variants that displayed reduced efficiency of plaque formation (C gene mutation, p23 and p28 or C-box sequence deletion, p19; Figure 5B) also showed slightly better efficiency of transformation (1.5–2.5-fold) in comparison to WT. The most substantial impairment in R-M system establishment (4-fold) was observed for the variant lacking a C gene (p30), which is consistent with the observation that it yielded the highest REase activity (Figure 5C). The *E. coli* cells carrying R-M system variants were also examined for cellular C protein levels (Figure 5D). Only the variants with a C gene mutation preventing C protein binding (p23 and p28) were undetectable using the western-blot assay. The C gene deletion mutant (p30) yielded similar results. Other variants showed C protein level comparable to WT (Figure 5D). These data together demonstrate that the presence of the C gene is essential for improving the efficiency of Csp231I plasmid transfer, but not required for REase production.

### C protein presence improves R-M system host fitness

We also questioned whether the WT R-M system and its C-deleted variant, which both have comparable restriction activity (p18 versus p30, Figure 5) would show any differences in viability or fitness. We challenged the strains in a direct one-flask competition assay by mixing equal numbers of cells carrying a WT R-M system or its C-deleted variant on a plasmid. Different antibiotic resistance genes served as the cell markers: p18amp<sup>WT</sup> (C+R+M+) vs. p30tet $\Delta$ C (C-R+M+) and in parallel controlled flasks restriction-negative R-M systems: p17amp (C+R-M+) versus p31tet (C-R-M+) were tested in identical conditions. Vector control (pBRamp versus pBRtet) served to normalize the experiments in case the antibiotics-resistance expression conferred some fitness advantage. We started the cultures and let them grow for about 130 generations in antibiotic-free minimal medium. Sub-cultures were created every 24 generations, and the number of colony-forming units were counted (Figure 6). The results show a significant gain in relative fitness advantage for WT R-M system over its C-deleted variant (WT/ $\Delta$ C, Figure 6) in which REase expression relies only on the separate tandem promoters. We also confirmed the competed cells maintained their high restrictive-phenotypes for the entire course of the experiment. For controls with restriction-negative R-M systems we found the difference in the relative fitness to be insubstantial, particularly, for those measurements taken within 100 generations (R- CWT versus R-  $\Delta$ C, Figure 6).

## DISCUSSION

We report here data on the regulation of expression for Csp231I R-M system with a C protein of a new class. According to the Sorokin classification C proteins fall into groups based on their distinct motifs in DNA sequence binding, which for C.Csp231I and its prototype C.EcoO109I (motif 8 – ACTAAGGA-T-TnCTTAGT) is



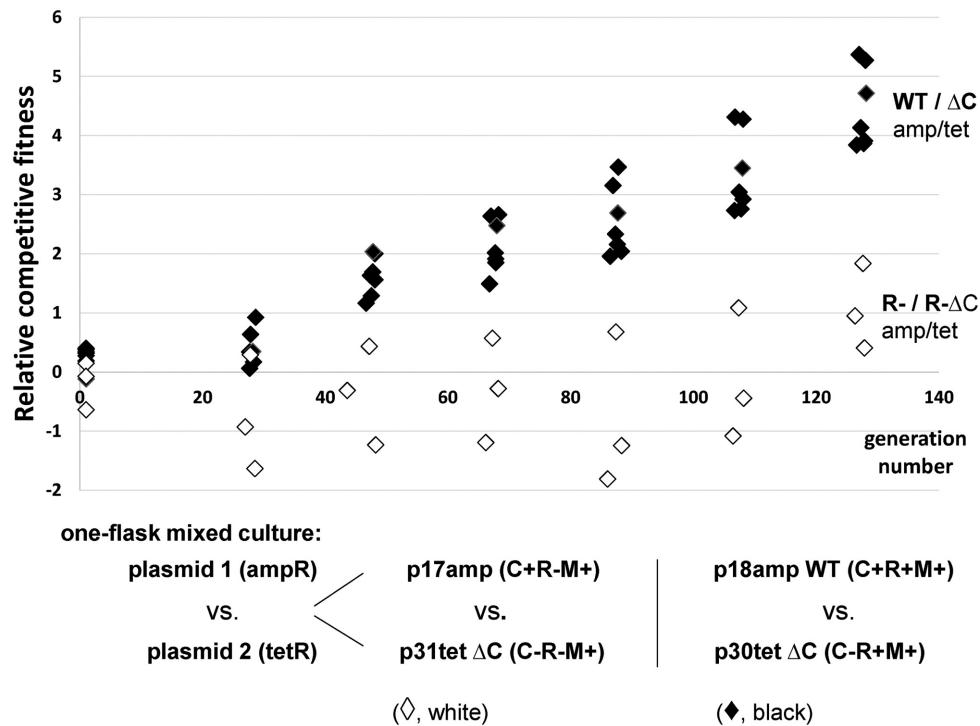
**Figure 5.** C protein is not essential for REase expression, but affects the R-M system transfer efficiency. Schematic diagram of the Csp231I R-M system WT and its variants (not in scale) are presented and data for particular panel A-D are adjusted to be read horizontally. The unchanged MTase gene is not shown. Changed elements are depicted in red color. Briefly: p18 (WT); p19 (deletion of  $O_L$  from C-box sequence); p23 (substitutions: A33G; R34E; Q37A in C protein); p28 (substitutions: S16A; Q17A; E18A in C protein); p30 (deletion of C gene and its upstream region including C-box and  $P_C$ ); p32 (mutation of -10 box of  $P_{R1}$ , TTAAAT  $\rightarrow$  CCCGGG); p24 (knock-out of REase gene, XhoI cut and Klenow filling) (A) Efficiency of plaque formation was assayed as plaque forming units on tested plasmid divided by plaque forming units on pBR322. The standard deviation from at least three experiments is shown. (B) Relative restriction is indicated as 1 for WT R-M system. (C) Efficiency of transformation is a relative number of transformants obtained from the same amount of plasmid DNAs carrying the indicated R-M system variants referred to vector pBR322 (D) Level of C protein in bacterial crude extracts was tested by western blot (horizontal bands) using the rabbit anti-C.Csp231I polyclonal antibodies and visualized by nitroblue tetrazolium (NBT) as the color development reagent. Culture samples were normalized by OD<sub>600</sub>. Loading control was prepared as one from Figure 3B; it is not shown for clarity. C.Csp231I protein is marked by arrow.

a small fraction of the entire C protein number (43,44). C.Csp231I and C.EcoO109I share about 70% identity in amino acid sequence and are significantly larger proteins at about 11 kDa (59). They are also different structurally from other studied C proteins with motifs 1–6 (~8–9 kDa). They possess two extra helices at the carboxyl terminus that may play a role in dimer formation as inferred from the solved C.Csp231I crystal structure (45,46). However, C.EcoO109I appears to operate by a mode of regulation typical for the majority of C proteins as its C gene inactivation leads to loss of REase production (33) and this is unlike the regulatory mode reported here for C.Csp231I.

#### Similarity in action of C.Csp231I to other R-M system control proteins

Despite the structural distinction of C.Csp231I and its C-box some essential features of transcriptional auto-control of other studied C proteins are maintained. We mapped the C gene promoter and determined its –10 and –35 hexamers. The produced transcript appears to be leaderless (70), which is a common property of operons for R-M systems associated with C proteins belonging to the C.PvuII and C.EcoRV families (27,35,40). The key element in C gene control over REase expression seems to be the shared mRNA present in majority of C-linked R-M systems as well as toxin-antitoxin systems associated with analogous tran-

scription factors (10,71,72). We also detected bicistronic mRNAs for the C and REase genes (Figure 4AB). The C-box of C.Csp231I is formed by two palindromic sequences CTAAG-n<sub>5</sub>-CTTAG separated by 18nt with A/T rich spacers (Figure 1C, bottom sequence). C binding to these sequences is likely to bend DNA tightly, as shown for other C proteins (33,37,68). In addition, C.Csp231I also uses an autoregulatory feed-back loop to switch the transcription profile from activation to repression. C proteins binding to the  $O_L$  site as a dimer (45) directly contact RNA polymerase  $\sigma^{70}$  subunits (region 4, R588) via a short highly conserved amino acid sequence (SQE) found in most C proteins (Figure 2A) and phage repressors (40,73). This contact results in transcription activation as seen for  $\lambda$ CI (69,74,75) and based on modeling the interaction between  $\sigma^{70}$  and in ternary complex with C.AhdI dimer and DNA (73). Indeed, a replacement of the key S<sub>16</sub>Q<sub>17</sub>E<sub>18</sub> residues of C.Csp231I led to loss of transcription activation as tested *in vivo* (Figure 3A). We expected that C binding for such variant would not be disturbed *in vitro*; however, binding of C-SQEmut variant to its WT C-box was not observed during our tests (Figure 2C). The analyzed C dimer – DNA crystal structure indicates the Q<sub>17</sub> residues are located close enough (<3 Å) to interact with the DNA phosphate backbone (Supplementary Figure S5; (46)). In addition, the S<sub>16</sub> substitution by A may destabilize the H2 helix as the ser-



**Figure 6.** Cells with the WT R-M system have a fitness advantage over cells carrying the C-deleted variant. Mixed cultures were prepared by adding equal numbers of the two competing *E. coli* strains into medium without any antibiotic (Materials and Methods). Each type carried plasmid with a specific R-M system variant and its distinct antibiotic marker (tetracycline or ampicillin), as indicated below the diagram. One flask co-cultures were diluted every 24 generations into fresh medium and CFUs of competing cells were measured. Relative competitive fitness ( $W$ ) was estimated individually for each mixed culture represented on diagram as a single symbol, calculated as  $W = \log(\text{CFU}_{\text{amp}}/\text{CFU}_{\text{tet}})$  and normalized to vector control (pBRamp versus pBRtet) (Materials and Methods). Black diamonds represents seven separate co-cultures, where WT R-M system in ampicillin resistant cells (p18amp) competed with C-deleted R-M system (p30tet) in tetracycline resistant cells. In control, parallel co-cultures, cells with plasmids with restriction-negative and modification-positive variants (p17amp vs. p31tet; white diamonds) were used.

ine residue exerts more propensity toward N-capping than the alanine residue (76). Overall, such an effect, through indirect readout, may subtly change the DNA's ability to achieve the required conformation as evidenced for some bacteriophage repressors (77,78). In a similar manner we tested the A<sub>33</sub>R<sub>34</sub>Q<sub>37</sub> replacement in C protein that is predicted to weaken DNA binding (45), and it showed comparable results to the S<sub>16</sub>Q<sub>17</sub>E<sub>18</sub> variant including loss of DNA binding *in vitro* (Figure 2C), inability to activate C-mediated activation of transcription (Figure 3A), decreased phage restriction *in vivo* (Figure 5AB) and diminished level of C protein in cells (Figure 5D). We also showed that as with other C proteins the repression step in transcription is associated with inverted repeats located closer to the C transcription start (O<sub>R</sub>), which are overlapped by the -35 hexamer of P<sub>C</sub> (Figure 3C). Saturation of O<sub>L</sub> and O<sub>R</sub> after C protein accumulation by a presumable C tetramer (46) leads to transcriptional autorepression. A similar switch in transcription (activation vs. repression) is linked to the highly cooperative binding of two C proteins dimers, and was identified for C.AhdI, C.PvuII, C.Esp1396I (36,38,79). Binding is not cooperative for C.Csp23II and C.EcoRV (27,46). Interestingly, the C.Csp23II activation achieved the highest values at C protein concentration below the detection threshold of our assay (Figure 3) suggesting that in this R-M system C protein regulation has been tuned to operate at low copy/chromosomal levels.

### Novel elements in R-M system regulation with C protein of a new class

The Csp23II R-M system contains collinearly oriented C and REase genes that are separated by 152 nt. We found that there are two functional promoters, P<sub>R1</sub> and P<sub>R2</sub>, for the REase gene in the region between the C and REase genes (Figure 1BC). This independent transcription is novel among the known C-associated R-M systems. A constitutive promoter for the REase was found in the Kpn2I R-M system, but its C gene is located divergently at some distance from the REase gene (39). The tandem promoters are the major source of REase transcription with only minor contribution of bicistronic mRNA driven by C gene promoter (Figure 4B). Our *in vivo* data show that the C gene effect on REase production is positive, but it is not large, which is unlike C.PvuII, where inactivation of C leads to complete loss of restriction activity (26). In this system abolishing C binding to DNA (SQE and ARQ variants) as tested *in vitro*, still maintains a restrictive phenotype (Figure 2C). The restriction activity was tested in plaque formation efficiency assays using  $\lambda$ vir phage (Figure 5BC), and was found to be decreased by 8–12-fold. This C protein positive effect is due to production of a shared C/R mRNA that is regulated by P<sub>C</sub>, and can be elevated in response C protein binding to its C-box (feed-back response) (Figure 4B, pLex-WT vs. pLex-Cmut). Deletion of the entire C gene resulted in a 3-fold

increase in phage restriction (Figure 5AB) and severely affected the host cell ability to compete with other restriction-positive cells (Figure 6). The reduction of REase expression in presence of the C gene may be the effect of transcription readthrough, which prevents RNA polymerase from binding to REase promoters. Hence REase expression is slightly attenuated in presence of C gene. Such modulation of toxic expression apparently is needed, otherwise leads to elevated toxicity and its host fitness compromised.

In this work, we report the first case of a C protein-associated R-M system in which the C gene positively affects REase production, but is not required for its expression. In general, the C protein control operates at two important stages: R-M system maintenance and establishment in a new cell (14). The maintenance stage was tested by the stability assay separately for WT and C-deleted R-M system and appeared to be comparable (data not shown). However, the direct competition fitness assay showed the cells carrying the C-deleted R-M system, in which REase expression relies on the separate tandem promoters, were outcompeted by the cells with WT R-M system. This result clearly indicates the C protein improves its host cell fitness, and is vital for R-M system propagation. Loss of fitness for C-deleted R-M system cells might occur due to autorestriction of its host genome when the R-M system activities are not finely balanced leading eventually to cell death and a heavily perturbed the mixed cell population ratio (8,9,80,81). We also tested whether the C.Csp231I may be essential when R-M system needs to enter the unprotected new cell. We demonstrated that the presence of the C gene significantly helps the R-M system during entry to a new cell (Figure 5C). The higher restriction activity the more lethal the effect appeared during the R-M system transfer. It remains to be determined how newly identified aspects of regulation mode for C.Csp231I and its R-M system exert the delay in expression of toxic restriction endonuclease in the cell. It seems such delay mechanism via C protein partial control, may not be sufficient for this particular system, and other patterns of temporal control may operate instead, e.g. by stimulation of MTase expression. The presence of the possible multi-layered complexity is not surprising as R-M systems and other toxin-antitoxin modules must be controlled to keep the counter-balancing amounts and timing properties to avoid lethality (10,12).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Kinga Ruszkiewicz and Anna Maliszewska for technical assistance. We thank also Robert Lintner, Bob Blumenthal, V. Nagaraja, Tadeusz Kaczorowski and Marian Sektas for reading the manuscript and interesting comments. We wish also to thank Stan Oldziej for help in C.Csp231I crystal analysis.

## FUNDING

National Center of Science (Poland) [7241/B/P01/2011/40 to I.M.]. Funding for open access charge: University of Gdansk, Poland.

*Conflict of interest statement.* None declared.

## REFERENCES

- Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
- Roberts, R.J., Vincze, T., Posfai, J. and Macelis, D. (2014) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.
- Vasu, K. and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.*, **77**, 53–72.
- Rocha, E.P., Danchin, A. and Viari, A. (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.*, **11**, 946–958.
- Oliveira, P.H., Touchon, M. and Rocha, E.P. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.*, **42**, 10618–10631.
- Loenen, W.A., Dryden, D.T., Raleigh, E.A., Wilson, G.G. and Murray, N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.*, **42**, 3–19.
- Pingoud, A., Wilson, G.G. and Wende, W. (2014) Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res.*, **42**, 7489–7527.
- Handa, N., Ichige, A., Kusano, K. and Kobayashi, I. (2000) Cellular responses to postsegregational killing by restriction-modification genes. *J. Bacteriol.*, **182**, 2218–2229.
- Katna, A., Boratynski, R., Furmanek-Blaszczak, B., Zolcinska, N. and Sektas, M. (2010) Unbalanced restriction impairs SOS-induced DNA repair effects. *J. Microbiol. Biotechnol.*, **20**, 30–38.
- Mruk, I. and Kobayashi, I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.*, **42**, 70–86.
- Naito, T., Kusano, K. and Kobayashi, I. (1995) Selfish behavior of restriction-modification systems. *Science*, **267**, 897–899.
- Hayes, F. and Kedzierska, B. (2014) Regulating toxin-antitoxin expression: controlled detonation of intracellular molecular time bombs. *Toxins (Basel)*, **6**, 337–358.
- Furuta, Y. and Kobayashi, I. (2012) Mobility of DNA sequence recognition domains in DNA methyltransferases suggests epigenetics-driven adaptive evolution. *Mobile Genet. Elem.*, **2**, 292–296.
- Mruk, I. and Blumenthal, R.M. (2008) Real-time kinetics of restriction-modification gene expression after entry into a new host cell. *Nucleic Acids Res.*, **36**, 2581–2593.
- Karyagina, A., Shilov, I., Tashlitskii, V., Khodoun, M., Vasil'ev, S., Lau, P.C. and Nikolskaya, I. (1997) Specific binding of sso II DNA methyltransferase to its promoter region provides the regulation of sso II restriction-modification gene expression. *Nucleic Acids Res.*, **25**, 2114–2120.
- Fedotova, E.A., Protsenko, A.S., Zakharova, M.V., Lavrova, N.V., Alekseevsky, A.V., Oretskaya, T.S., Karyagina, A.S., Solonin, A.S. and Kubareva, E.A. (2009) SsoII-like DNA-methyltransferase Ecl18kI: interaction between regulatory and methylating functions. *Biochemistry (Mosc.)*, **74**, 85–91.
- Butler, D. and Fitzgerald, G.F. (2001) Transcriptional analysis and regulation of expression of the ScrFI restriction-modification system of *Lactococcus lactis* subsp. *cremoris* UC503. *J. Bacteriol.*, **183**, 4668–4673.
- Beletskaya, I.V., Zakharova, M.V., Shlyapnikov, M.G., Semenova, L.M. and Solonin, A.S. (2000) DNA methylation at the CfrBI site is involved in expression control in the CfrBI restriction-modification system. *Nucleic Acids Res.*, **28**, 3817–3822.
- Christensen, L.L. and Josephsen, J. (2004) The methyltransferase from the LlaDII restriction-modification system influences the level of expression of its own gene. *J. Bacteriol.*, **186**, 287–295.

20. Protsenko, A., Zakharova, M., Nagornykh, M., Solonin, A. and Severinov, K. (2009) Transcription regulation of restriction-modification system Ecl18kI. *Nucleic Acids Res.*, **37**, 5322–5330.
21. Liu, Y., Ichige, A. and Kobayashi, I. (2007) Regulation of the EcoRI restriction-modification system: Identification of *ecoRIM* gene promoters and their upstream negative regulators in the *ecoRIR* gene. *Gene*, **400**, 140–149.
22. Liu, Y. and Kobayashi, I. (2007) Negative regulation of EcoRI restriction enzyme gene associated with intragenic reverse promoters. *J. Bacteriol.*, **189**, 6928–6935.
23. Mruk, I., Liu, Y., Ge, L. and Kobayashi, I. (2011) Antisense RNA associated with biological regulation of a restriction-modification system. *Nucleic Acids Res.*, **39**, 5622–5632.
24. Nagornykh, M., Zakharova, M., Protsenko, A., Bogdanova, E., Solonin, A.S. and Severinov, K. (2011) Regulation of gene expression in restriction-modification system Eco29kI. *Nucleic Acids Res.*, **39**, 4653–4663.
25. Ives, C.L., Nathan, P.D. and Brooks, J.E. (1992) Regulation of the BamHI restriction-modification system by a small intergenic open reading frame, *bamHIC*, in both *Escherichia coli* and *Bacillus subtilis*. *J. Bacteriol.*, **174**, 7194–7201.
26. Tao, T., Bourne, J.C. and Blumenthal, R.M. (1991) A family of regulatory genes associated with type II restriction-modification systems. *J. Bacteriol.*, **173**, 1367–1375.
27. Semenova, E., Minakhin, L., Bogdanova, E., Nagornykh, M., Vasilov, A., Heyduk, T., Solonin, A., Zakharova, M. and Severinov, K. (2005) Transcription regulation of the EcoRV restriction-modification system. *Nucleic Acids Res.*, **33**, 6942–6951.
28. Ives, C.L., Sohail, A. and Brooks, J.E. (1995) The regulatory C proteins from different restriction-modification systems can cross-complement. *J. Bacteriol.*, **177**, 6313–6315.
29. Vijesurier, R.M., Carlock, L., Blumenthal, R.M. and Dunbar, J.C. (2000) Role and mechanism of action of C.PvuII, a regulatory protein conserved among restriction-modification systems. *J. Bacteriol.*, **182**, 477–487.
30. Anton, B.P., Heiter, D.F., Benner, J.S., Hess, E.J., Greenough, L., Moran, L.S., Slatko, B.E. and Brooks, J.E. (1997) Cloning and characterization of the BglII restriction-modification system reveals a possible evolutionary footprint. *Gene*, **187**, 19–27.
31. Cesnaviciene, E., Mitkaite, G., Stankevicius, K., Janulaitis, A. and Lubys, A. (2003) Esp1396I restriction-modification system: structural organization and mode of regulation. *Nucleic Acids Res.*, **31**, 743–749.
32. Kroger, M., Blum, E., Deppe, E., Dusterhoff, A., Erdmann, D., Kilz, S., Meyer-Rogge, S. and Mostl, D. (1995) Organization and gene expression within restriction-modification systems of *Herpetosiphon giganteus*. *Gene*, **157**, 43–47.
33. Kita, K., Tsuda, J. and Nakai, S.Y. (2002) C.EcoO109I, a regulatory protein for production of EcoO109I restriction endonuclease, specifically binds to and bends DNA upstream of its translational start site. *Nucleic Acids Res.*, **30**, 3558–3565.
34. Rimseliene, R., Vaisvila, R. and Janulaitis, A. (1995) The *eco72IC* gene specifies a trans-acting factor which influences expression of both DNA methyltransferase and endonuclease from the Eco72I restriction-modification system. *Gene*, **157**, 217–219.
35. Bogdanova, E., Djordjevic, M., Papapanagiotou, I., Heyduk, T., Kneale, G. and Severinov, K. (2008) Transcription regulation of the type II restriction-modification system AhdI. *Nucleic Acids Res.*, **36**, 1429–1442.
36. Mruk, I., Rajesh, P. and Blumenthal, R.M. (2007) Regulatory circuit based on autogenous activation-repression: roles of C-boxes and spacer sequences in control of the PvuII restriction-modification system. *Nucleic Acids Res.*, **35**, 6935–6952.
37. Mruk, I. and Blumenthal, R.M. (2009) Tuning the relative affinities for activating and repressing operators of a temporally regulated restriction-modification system. *Nucleic Acids Res.*, **37**, 983–998.
38. Bogdanova, E., Zakharova, M., Streeter, S., Taylor, J., Heyduk, T., Kneale, G. and Severinov, K. (2009) Transcription regulation of restriction-modification system Esp1396I. *Nucleic Acids Res.*, **37**, 3354–3366.
39. Lubys, A., Jurenaite, S. and Janulaitis, A. (1999) Structural organization and regulation of the plasmid-borne type II restriction-modification system Kpn2I from *Klebsiella pneumoniae* RFL2. *Nucleic Acids Res.*, **27**, 4228–4234.
40. Knowle, D., Lintner, R.E., Touma, Y.M. and Blumenthal, R.M. (2005) Nature of the promoter activated by C.PvuII, an unusual regulatory protein conserved among restriction-modification systems. *J. Bacteriol.*, **187**, 488–497.
41. Williams, K., Savageau, M.A. and Blumenthal, R.M. (2013) A bistable hysteretic switch in an activator-repressor regulated restriction-modification system. *Nucleic Acids Res.*, **41**, 6045–6057.
42. Liang, J. and Blumenthal, R.M. (2013) Naturally-occurring, dually-functional fusions between restriction endonucleases and regulatory proteins. *BMC Evol. Biol.*, **13**, 218.
43. Sorokin, V., Severinov, K. and Gelfand, M.S. (2009) Systematic prediction of control proteins and their DNA binding sites. *Nucleic Acids Res.*, **37**, 441–451.
44. Sorokin, V., Severinov, K. and Gelfand, M.S. (2010) Large-scale identification and analysis of C-proteins. *Methods Mol. Biol.*, **674**, 269–282.
45. McGeehan, J.E., Streeter, S.D., Thresh, S.J., Taylor, J.E., Shevtsov, M.B. and Kneale, G.G. (2011) Structural analysis of a novel class of R-M controller proteins: C.Csp231I from *Citrobacter* sp. RFL231. *J. Mol. Biol.*, **409**, 177–188.
46. Shevtsov, M.B., Streeter, S.D., Thresh, S.J., Swiderska, A., McGeehan, J.E. and Kneale, G.G. (2015) Structural analysis of DNA binding by C.Csp231I, a member of a novel class of R-M controller proteins regulating gene expression. *Acta Crystallogr. D Biol. Crystallogr.*, **71**, 398–407.
47. Mruk, I. and Kaczorowski, T. (2003) Genetic organization and molecular analysis of the EcoVIII restriction-modification system of *Escherichia coli* E1585–68 and its comparison with isospecific homologs. *Appl. Environ. Microbiol.*, **69**, 2638–2650.
48. Mruk, I., Sektas, M. and Kaczorowski, T. (2001) Characterization of pEC156, a ColE1-type plasmid from *Escherichia coli* E1585–68 that carries genes of the EcoVIII restriction-modification system. *Plasmid*, **46**, 128–139.
49. Casadaban, M.J. and Cohen, S.N. (1980) Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. *J. Mol. Biol.*, **138**, 179–207.
50. Novy, R., Drott, D., Yaeger, K. and Mierendorf, R. (2001) Overcoming the codon bias of *E. coli* for enhanced protein expression. *InNovations*, **12**, 1–2.
51. Guzman, L.M., Belin, D., Carson, M.J. and Beckwith, J. (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.*, **177**, 4121–4130.
52. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1986) *Molecular Cloning: A Laboratory Manual*. 2nd edn. Cold Spring Harbor Laboratory Press, NY.
53. Platko, J.V., Willins, D.A. and Calvo, J.M. (1990) The *ilvIH* operon of *Escherichia coli* is positively regulated. *J. Bacteriol.*, **172**, 4563–4570.
54. Simons, R.W., Houman, F. and Kleckner, N. (1987) Improved single and multicopy lac-based cloning vectors for protein and operon fusions. *Gene*, **53**, 85–96.
55. Diederich, L., Roth, A. and Messer, W. (1994) A versatile plasmid vector system for the regulated expression of genes in *Escherichia coli*. *Biotechniques*, **16**, 916–923.
56. Schagger, H. and von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal. Biochem.*, **166**, 368–379.
57. Lane, D. and Harlow, E. (1988) *Antibodies: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY.
58. Leroi, A.M., Bennett, A.F. and Lenski, R.E. (1994) Temperature acclimation and competitive fitness: an experimental test of the beneficial acclimation assumption. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 1917–1921.
59. Mruk, I. and Kaczorowski, T. (2007) A rapid and efficient method for cloning genes of type II restriction-modification systems by use of a killer plasmid. *Appl. Environ. Microbiol.*, **73**, 4286–4293.
60. Nwankwo, D.O., Moran, L.S., Slatko, B.E., Waite-Rees, P.A., Dorner, L.F., Benner, J.S. and Wilson, G.G. (1994) Cloning, analysis and expression of the HindIII R-M-encoding genes. *Gene*, **150**, 75–80.
61. Werbowy, O., Boratynski, R., Dekowska, A. and Kaczorowski, T. (2014) Genetic analysis of maintenance of pEC156, a naturally occurring *Escherichia coli* plasmid that carries genes of the EcoVIII restriction-modification system. *Plasmid*, **77**, 39–50.

62. Madsen, A. and Josephsen, J. (1998) Cloning and characterization of the lactococcal plasmid-encoded type II restriction/modification system, LlaDII. *Appl. Environ. Microbiol.*, **64**, 2424–2431.
63. Madsen, A. and Josephsen, J. (1998) Characterization of LlaCI, a new restriction-modification system from *Lactococcus lactis* subsp. *cremoris* W15. *Biol. Chem.*, **379**, 443–449.
64. Kaw, M.K. and Blumenthal, R.M. (2010) Translational independence between overlapping genes for a restriction endonuclease and its transcriptional regulator. *BMC Mol. Biol.*, **11**, 87.
65. Shultzaberger, R.K., Chen, Z., Lewis, K.A. and Schneider, T.D. (2007) Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Res.*, **35**, 771–788.
66. Han, X. and Turnbough, C.L. Jr (2014) Transcription start site sequence and spacing between the -10 region and the start site affect reiterative transcription-mediated regulation of gene expression in *Escherichia coli*. *J. Bacteriol.*, **196**, 2912–2920.
67. Turnbough, C.L. Jr (2011) Regulation of gene expression by reiterative transcription. *Curr. Opin. Microbiol.*, **14**, 142–147.
68. McGeehan, J.E., Streeter, S.D., Thresh, S.-J., Ball, N., Ravelli, R.B.G. and Kneale, G.G. (2008) Structural analysis of the genetic switch that regulates the expression of restriction-modification genes. *Nucleic Acid Res.*, **36**, 4778–4787.
69. Jain, D., Nickels, B.E., Sun, L., Hochschild, A. and Darst, S.A. (2004) Structure of a ternary transcription activation complex. *Mol. Cell*, **13**, 45–53.
70. Moll, I., Grill, S., Gualerzi, C.O. and Blasi, U. (2002) Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol. Microbiol.*, **43**, 239–246.
71. de la Hoz, A.B., Ayora, S., Sitkiewicz, I., Fernandez, S., Pankiewicz, R., Alonso, J.C. and Ceglowski, P. (2000) Plasmid copy-number control and better-than-random segregation genes of pSM19035 share a common regulator. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 728–733.
72. Hallez, R., Geeraerts, D., Sterckx, Y., Mine, N., Loris, R. and Van Melderen, L. (2010) New toxins homologous to ParE belonging to three-component toxin-antitoxin systems in *Escherichia coli* O157:H7. *Mol. Microbiol.*, **76**, 719–732.
73. McGeehan, J.E., Streeter, S.D., Papapanagiotou, I., Fox, G.C. and Kneale, G.G. (2005) High-resolution crystal structure of the restriction-modification controller protein C.AhdI from *Aeromonas hydrophila*. *J. Mol. Biol.*, **346**, 689–701.
74. Kolkhof, P. and Muller-Hill, B. (1994) Lambda cI repressor mutants altered in transcriptional activation. *J. Mol. Biol.*, **242**, 23–36.
75. Bushman, F.D., Shang, C. and Ptashne, M. (1989) A single glutamic acid residue plays a key role in the transcriptional activation function of lambda repressor. *Cell*, **58**, 1163–1171.
76. Doig, A.J., Chakrabarty, A., Klingler, T.M. and Baldwin, R.L. (1994) Determination of free energies of N-capping in alpha-helices by modification of the Lifson-Roig helix-coil theory to include N- and C-capping. *Biochemistry*, **33**, 3396–3403.
77. Koudelka, G.B., Mauro, S.A. and Ciubotaru, M. (2006) Indirect readout of DNA sequence by proteins: the roles of DNA sequence-dependent intrinsic and extrinsic forces. *Prog. Nucleic Acids Res. Mol. Biol.*, **81**, 143–177.
78. Mauro, S.A., Pawlowski, D. and Koudelka, G.B. (2003) The role of the minor groove substituents in indirect readout of DNA sequence by 434 repressor. *J. Biol. Chem.*, **278**, 12955–12960.
79. McGeehan, J.E., Papapanagiotou, I., Streeter, S.D. and Kneale, G.G. (2006) Cooperative binding of the C.AhdI controller protein to the C/R promoter and its role in endonuclease gene expression. *J. Mol. Biol.*, **358**, 523–531.
80. Kobayashi, I. (2004) Restriction-modification systems as minimal forms of life. In: Pingoud, A. (ed). *Restriction endonucleases*. Springer-Verlag, Berlin, pp. 19–62.
81. Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N. and Uchiyama, I. (1999) Shaping the genome—restriction-modification systems as mobile genetic elements. *Curr. Opin. Genet. Dev.*, **9**, 649–656.