6982–6995 *Nucleic Acids Research, 2021, Vol. 49, No. 12* *Published online 23 June 2021*
https://doi.org/10.1093/nar/gkab524

$TnpA_{REP}$ and REP sequences dissemination in bacterial genomes: REP recognition determinants

Alix Corneloup[1], Anne Caumont-Sarcos[1], Alain Kamgoue[2], Brigitte Marty[1], Phan Thai Nguyen Le[1], Patricia Siguier[1], Catherine Guynet[1,*] and Bao Ton-Hoang[1,*]

[1]Laboratoire de Microbiologie et de Génétique Moléculaires (LMGM), CBI, CNRS, Université Toulouse UPS, Toulouse, France and [2]Image Processing Facility, CBI, Toulouse, France

Received September 25, 2020; Revised May 27, 2021; Editorial Decision May 29, 2021; Accepted June 17, 2021

ABSTRACT

REP, diverse palindromic DNA sequences found at high copy number in many bacterial genomes, have been attributed important roles in cell physiology but their dissemination mechanisms are poorly understood. They might represent non-autonomous transposable elements mobilizable by $TnpA_{REP}$, the first prokaryotic domesticated transposase associated with REP. $TnpA_{REP}$, fundamentally different from classical transposases, are members of the HuH superfamily and closely related to the transposases of the IS*200*/IS*605* family. We previously showed that *Escherichia coli* $TnpA_{REP}$ processes cognate single stranded REP *in vitro* and that this activity requires the integrity of the REP structure, in particular imperfect palindromes interrupted by a bulge and preceded by a conserved DNA motif. A second group of REPs rather carry perfect palindromes, raising questions about how the latter are recognized by their cognate $TnpA_{REP}$. To get insight into the importance of REP structural and sequence determinants in these two groups, we developed an *in vitro* activity assay coupled to a mutational analysis for three different $TnpA_{REP}$/REP duos via a SELEX approach. We also tackled the question of how the cleavage site is selected. This study revealed that two $TnpA_{REP}$ groups have co-evolved with their cognate REPs and use different strategies to recognize their REP substrates.

INTRODUCTION

Although bacterial genomes are small and compact compared to their eukaryotic counterparts, they harbor multiple repeated sequences playing various functions (for review see (1,2)). Among them, REP elements (for Repetitive Extragenic Palindrome) are small palindromic sequences of 20–50 nts preceded by a conserved tetranucleotide, most often GTAG. REPs are present in great numbers, mostly in intergenic regions of bacterial genomes: about six hundred in the *Escherichia coli* K12 genome or thousands of copies in some *Pseudomonas* strains. They are often organized in BIMEs (for Bacterial Interspersed Multiple Elements). These structures combine two REPs in inverse orientation, REP and inverted REP (iREP), separated by a variable linker and frequently found as consecutive tandem copies. Various cellular functions have been attributed to REP/BIME in the structuring and plasticity of the genome, or in the regulation of gene expression at transcriptional, post-transcriptional levels, and in the regulation of stress response (3–10).

A $tnpA_{REP}$ gene was described to be associated with REPs (11) in its immediate proximity in structures called REPtrons (23) (see examples in Figure 1). For simplicity later on in the text, we will refer to REPtron as to a given encoded protein $TnpA_{REP}$ and the ensemble of cognate REPs. It is important to note that the majority of REPs are generally distributed genome-wide but a given $tnpA_{REP}$ exists in most cases as a single copy and there is no evidence of $tnpA_{REP}$ mobility. While the presence of $tnpA_{REP}$ is often found to be correlated with the abundance of REPs in a given genome (11,12), $tnpA_{REP}$ behavior, based on several criteria (copy number per replicon, presence on plasmids, duplication rates) resembles more housekeeping genes than transposase genes (13). $TnpA_{REP}$ has thus been proposed to be a domesticated transposase mobilizing REPs over bacterial genomes. However, the underlying dissemination mechanism remains to be elucidated.

$TnpA_{REP}$ are members of the HuH recombinase superfamily, which includes Rep proteins (rolling circle replication RCR, not to be confused with REP), relaxases (conjugative transfer) and certain Transposases (Helitrons, IS*91*/IS*CR* and IS*200*/IS*605* families). All these proteins cleave, join DNA and carry the characteristic HuH motif (histidine-hydrophobic residue-histidine) crucial for coordinating a metal ion. The metal ion is essential for the nucleophilic attack by the characteristic catalytic Tyr residue, generating a covalent 5′ P-tyrosine intermediate and a free

*To whom correspondence should be addressed. Email: bao.ton-hoang@univ-tlse3.fr
Correspondence may also be addressed to Catherine Guynet. Email: catherine.guynet@univ-tlse3.fr

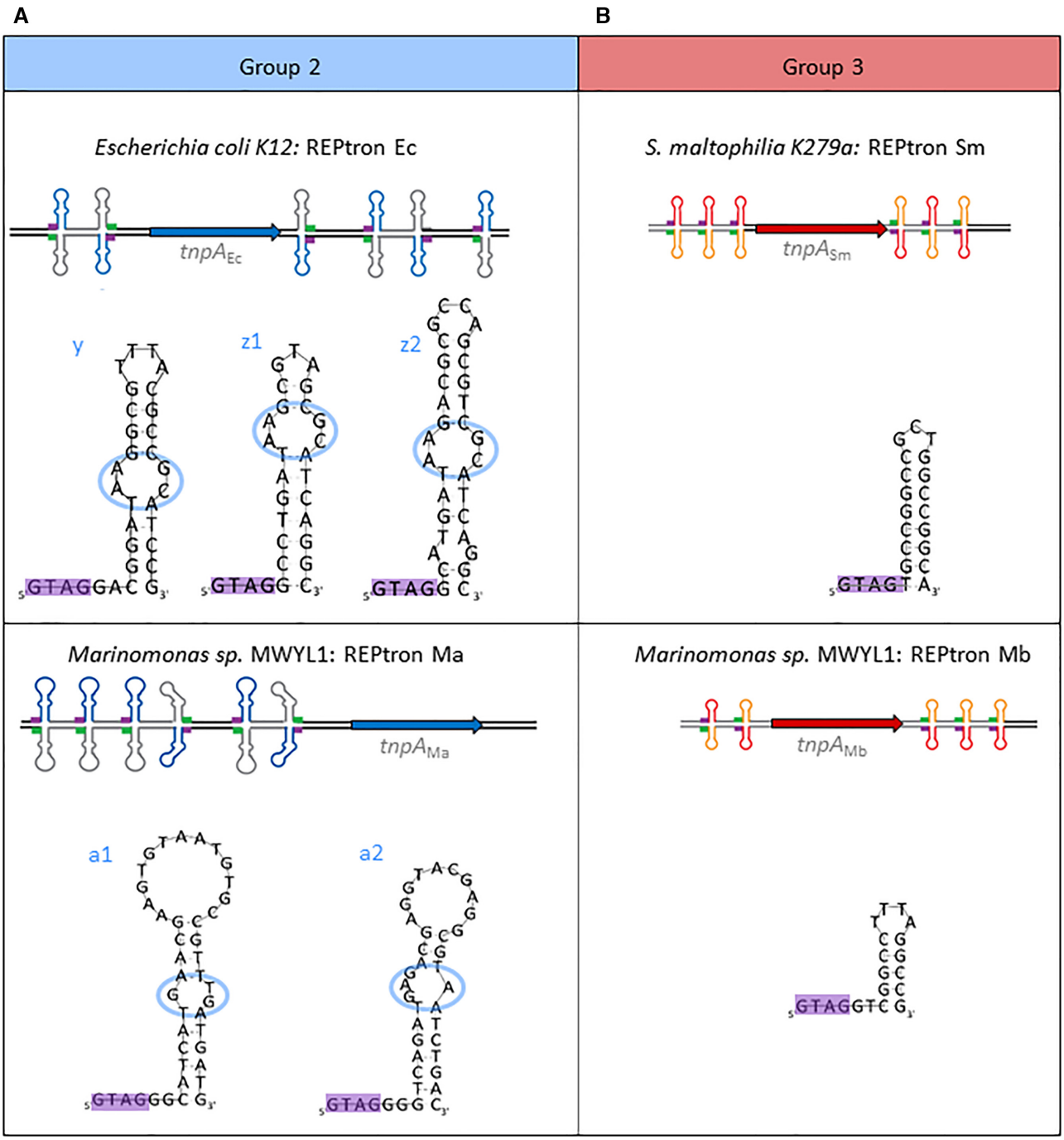

Figure 1. Model REPtrons and corresponding $TnpA_{REP}$/REPs. (**A**) Group 2: *E. coli* MG1655 REPtron Ec, the principal model of the group 2 and three classes of REP y, z1, z2 (top). Analysis of this group was complemented with *in vitro* assays of $TnpA_{Ec}$ on REP_{Ma1} and REP_{Ma2} originated from *Marinomonas* sp. MWYL1 REPtron Ma (bottom). (**B**) Group 3: *S. maltophilia* K279a REPtron Sm (top) and *Marinomonas* sp. MWYL1 REPtron Mb (bottom). In the presented REPtrons, $tnpA_{REP}$ (bold arrow) is represented in blue and red in group 2 and 3, respectively. REP and iREP structures are represented in blue/grey (group 2) and red/orange (group 3), respectively, purple and green bold lines—GTAG motif and complementary sequence CTAC, respectively. In the REP detailed structures, the GTAG is boxed in purple. Blue ovals represent irregularities in the group 2 REPs stem. This colour code is maintained throughout the text. For simplicity, in REPtron Ec, y, z1 and z2 REPs are presented without distinction.

3′OH after DNA cleavage. Then, the latter 3′OH extremity can serve as primer for RCR, or act as nucleophile to attack the P-tyrosine bond to resolve it (for more details see (14)). $TnpA_{REP}$, while constituting a separate family, are closely related to the transposases of the IS*200*/IS*605* family ($TnpA_{IS200/IS605}$) of bacterial insertion sequences (IS), which members are bordered by palindromic ends (for review see (15)). Transposition of IS*200*/IS*605* elements occurs on single strand (ss) DNA and is strand-specific (16–18). Moreover, IS*200*/IS*605* cleavage sites are chosen via a peculiar DNA-DNA complementarity between the cleavage sites and the respective 'guide' sequences located 5′ to each palindrome (19), (example of model element IS*608* in Supplementary Figure S1).

tnpA $_{REP}$ has been found in about 25% of all bacterial species (13). They are present largely in γ-proteobacteria, but also exist in other distant genera. Based on their protein sequences, $TnpA_{REP}$ can be classified into several groups. In particular, groups 2 and 3 (13) (also called groups 2.2 and 2.5, respectively (20)) are associated with the first and best described REPs (7,11,12,21,22). Group 2 mostly includes $TnpA_{REP}$ from different enterobacteria, while group 3 mainly comprises members from *Pseudomonas* species.

These two $TnpA_{REP}$ groups are associated with two types of REPs. Group 2 $TnpA_{REP}$ are associated with long REPs interrupted by an irregular zone/bulge in their stems (Figure 1A, Supplementary Figure S2A top), while group 3 REPs are short and generally perfectly palindromic (Figure 1B, Supplementary Figure S2A bottom). The group 2 $TnpA_{REP}$ from *E. coli* ($TnpA_{Ec}$) is the sole $TnpA_{REP}$ for which experimental studies of interactions with REP substrates have been performed. We have previously shown that $TnpA_{Ec}$ specifically recognizes ss REP (but not iREP) and catalyzes its cleavage and recombination *in vitro*. Cleavage occurs at the dinucleotide CT situated 5′ or 3′ to the REP structure (23). The conserved tetranucleotide GTAG is crucial for this activity. Consistent with this functional role, the GTAG motif forms contacts with several TnpAREP residues, as shown in the co-crystal (24) (see Figure 6B bottom). *E. coli* REPs (REP_{Ec}) include two conserved mismatches that form a bulge within the REP stem (Figure 1A). This bulge is required for activity since compensatory mutations restoring regular stem eliminated activity. Although these analyses helped to shed light on the importance of the conserved tetranucleotide GTAG and the bulge in REP_{Ec} recognition by $TnpA_{Ec}$, the role of other components (loop, stem) was still ambiguous. How group 3 $TnpA_{REP}$ recognize their perfect palindromic REPs as well as how the cleavage site is selected remain to be elucidated.

Here, to go further in deciphering $TnpA_{REP}$ activity, we developed a sensitive *in vitro* activity assay, CST (for Cleavage and Strand Transfer) to detect and map REP cleavage sites, that we then adapted to a CST-based SELEX. A combination of this robust approach with a mutational analysis permitted to re-examine and to get access to the importance of different structural features in REP recognition by group 2 $TnpA_{REP}$. In parallel, we extended the analysis to the group 3, for which no data are available, and tackled the question of cleavage site selection in this group. We showed that each group uses different strategies to recognize its REP substrates and demonstrate the role of the GTAG motif in cleavage site selection for a group 3 member. These results represent considerable progress in the comprehension of the distinct mechanism of $TnpA_{REP}$ mediated mobility and specificity of these expanding elements, which led us to discuss REPtrons potential evolutionary routes.

MATERIALS AND METHODS

$TnpA_{REP}$ purification

$TnpA_{Ec}$-His6 was purified as previously described (23). $TnpA_{Mb}$ and $TnpA_{Sm}$ coding sequences were synthetized and cloned in suitable expression vector under control of arabinose promotor. $TnpA_{Mb}$ and $TnpA_{Sm}$ were purified by affinity as N-term STREP tag fusion proteins, corresponding proteins were expressed in the *E. coli* K12 Strain Rosetta (DE3) (Novagen). A preculture was grown at 37°C in L broth containing Amp was diluted 50-fold into the same medium at 30°C. Protein expression was induced at OD_{600} = 0.5–0.6 by adding arabinose to 0.8% final. After 3h, bacteria were centrifuged and the pellet was resuspended in buffer NP (phosphate buffer (NaH_2PO_4 and Na_2HPO_4) pH 8 50 mM, NaCl 400 mM, Triton 0.2%, glycerol 10%, DTT 1 mM) supplemented with 1 mg/ml lysozyme and EDTA-free protease inhibitor cocktail (Roche). Bacteria were sonicated and the lysate was cleared by centrifugation. The supernatant was then mixed with resin Strep-tactin Superflow Plus (Qiagen) during 2h at 4°C. After washes in buffer NP, the proteins were eluted in buffer NPD (phosphate buffer (NaH_2PO_4 and Na_2HPO_4) pH 8 50 mM, NaCl 400 mM, Triton 0.2%, glycerol 10%, DTT 1 mM, desthiobiotine 2.5 mM). An additional purification step was performed using a Superdex 200 column (Highload 16/60, GE Healthcare). The samples were then dialysed in 25 mM HEPES pH 7.5, 400 mM NaCl, 1 mM EDTA, 1 mM DTT and 20% glycerol and stored at –80°C.

Standard reactions *in vitro*

Oligonucleotides (Eurofins Genomics) were 5′-end-labelled with [γ-^{32}P] ATP (Perkin Elmer) using T4 polynucleotide kinase (Thermo scientific). Labelled oligonucleotides were purified on a G25 column (GE Healthcare).

0.02 μM 5′-labelled oligonucleotide and 0.5 μM unlabelled oligonucleotide were incubated with $TnpA_{REP}$ (45 min, 37°C, final volume 10 μl) in 12.5 mM Tris (pH 7.5), 120 mM NaCl, 5 mM $MnCl_2$/$MgCl_2$, 1 mM DTT, 20 μg/ml BSA, 0.5 μg of poly-dIdC and 7% glycerol. Reactions were separated on an 8% denaturing gel (7 M urea, Tris Borate EDTA 2 mM, acrylamide 19:1 8%, migration at room temperature at 50 W) and analysed by phosphorimaging. In EMSA experiments, labelled substrates were incubated with corresponding $TnpA_{REP}$ in reaction buffer without divalent metal cation and complexes were separated on 8% native acrylamide gel (Tris acetate EDTA, acrylamide 37.5:1 8%, glycerol 7%, migration at 10 V/cm, 4°C) and analysed by phosphorimaging.

CST- test on circular substrates *in vitro*

Proteins and substrates were incubated together 45 min at 37°C in the reaction buffer in a final volume of 10 μl

containing 50 ng of ∼4kb pBluescript SK- derivative ss phagemid circular substrate, 0.5 μg of poly-dIdC, 1.5 μM $TnpA_{REP}$. 3 μl of 10 μM stock of attacking oligonucleotide B457 were added and incubation continued for 30 min. Reaction was stopped and de-proteinized by adding an equal volume of 25 mM EDTA, 0,6 mg/ml Proteinase K and 2% SDS and incubated for 1 h at 37°C. Products were purified on Promega columns (Wizard SV Gel and PCR) and subsequently served as templates for PCR amplification with GoTaq polymerase using B457 and Cy5 or Cy3 substrate specific fluorescent primer (98°C, 2 min, 30× (98°C 30 s, 56°C 30 s, 72°C 30 s)). PCR products were separated on a 8% native polyacrylamide gel and revealed by scan on GE Healthcare Typhoon Trio Imager.

CST-based selex

1 μl of 1 μM of degenerate substrates (Eurofins Genomics) was incubated with the corresponding $TnpA_{REP}$ in the standard reactional mixture for 45 min at 37°C. The following steps were as described for CST. Amplification was carried out with 457 or other attacking primers and 321, common for all substrates. After sequencing with 321, ss substrates were prepared for next round by asymetric PCR with Phusion polymerase using 0.1 μM 321 and 10 μM of corresponding attacking primer (98°C 30 s, 45× (98°C 10 s, 56°C 10 s, 72°C 10 s)). The quantification procedure is described in details in Supplementary Materials and in Supplementary Figure S3C.

RESULTS

Experimental REPtron models

In this study, we focused on *E. coli* MG1655 REPtron (called Ec) as principal model for the group 2 (Figure 1A, top). *E. coli* MG1655 genome harbors 3 types of REP: y (35nts), z1 (29nts) and z2 (37nts) often combined in BIME as mosaics of y-z1 or y-z2 REPs at multiple loci in the genome (7). The three REP_{Ec} types are imperfect palindromes preceded by the characteristic GTAG and can form stem-loop structures interrupted by a conserved AA-GC mismatch forming a bulge, and certain unpaired bases in the loop. In addition, they share several conserved positions in the stem (Supplementary Figure S2B).

To investigate the group 3 $TnpA_{REP}$/REP, several $TnpA_{REP}$ candidates were tested for their expression and solubility in *E. coli*. We chose *Stenotrophomonas maltophilia* K279a and *Marinomonas sp*. MWYL1 genomes (Figure 1B). Organisms of this group often host several REPtrons and carry hundreds of REPs in their genomes (12). Furthermore, *Stenostrephomonads* are omnipresent environmental bacteria often present in the soil, and *S. maltophilia* is an opportunistic pathogen commonly associated with hospital acquired infections. A phylogenetic analysis of REP distribution in a *S. maltophilia* collection has pointed out a dynamic character of the REP/BIME distribution in these genomes suggesting an ongoing proliferation process (12). We chose to study Sm, one of REPtrons in the *S. maltophilia* K279a strains. REPtron Sm carries perfect palindromes REP (REP_{Sm}) of 16 nts interrupted by 3 nts and directly preceded by the conserved tetranucleotide GTAG (Figure 1B).

Marinomonas sp. MWYL1 genome carries a group 3 REPtron Mb (Figure 1B, bottom) and also a group 2 REPtron Ma (Figure 1A, bottom and see below). REPtron Mb comprises small perfect palindromic REPs (REP_{Mb}) of 10 nts, interrupted by 4 nts and separated by 2 bases to the GTAG tetranucleotide. Interestingly, in contrast to the general genome-wide distribution, for both REPtron Ma and REPtron Mb, a physical association between $tnpA_{REP}$ genes and REPs is quite pronounced (11). REP_{Ma} and REP_{Mb} are concentrated in proximity to $tnpA_{Ma}$ and $tnpA_{Mb}$, suggesting that the arrival of these REPtrons was recent and that the corresponding REP copies have been subsequently multiplied in their vicinity.

We concentrated our analyses principally on REPtrons Ec, Sm and Mb. The three purified $TnpA_{REP}$ were then used to examine their activities on their cognate REPs. The study on the group 2 was also supplemented by activity tests performed with $TnpA_{Ec}$ on group 2 REP_{Ma} from *Marinomonas sp*. MWYL1 genome (Figure 1A, bottom). REPtron Ma includes two types of long REPs (REP_{Ma1} and REP_{Ma2} of 42 and 38 nts) with different irregularities in the stems followed by large loops for which an alignment showed few conserved positions (Supplementary Figure S2C).

Cleavage and strand transfer assay (CST)

We previously showed that $TnpA_{Ec}$ is capable of cleaving and recombining ss REP_{Ec} *in vitro* (23). Cleavages occur 5′ or 3′ of REP substrates at a dinucleotide C|T. To go further in the comprehension of REP mobility mechanism, we developed an activity assay called CST (Cleavage-Strand Transfer). The CST assay takes advantage of the general property of HuH enzymes, which form a 5′P-tyrosine link and a 3′OH extremity upon cleavage (14) (Figure 2A3). The 3′-OH then can be differently used. Upon cleavage by Rep proteins (single-stranded phages and RCR plasmids) and conjugative relaxases, the 3′-OH group can serve to prime replication. The 3′-OH can also act as the nucleophile for strand transfer to resolve the 5′P-tyrosine link in the termination step of RCR replication, conjugative transfer and transposition. Both possibilities might be exploited to disseminate REP/BIME sequences (23).

The CST assay was first developed with the REPtron Ec (Figure 2). After incubation of ss REP substrates with $TnpA_{Ec}$ in a reaction buffer allowing cleavages to occur (Figure 2A2-3), an excess of an 'attacking' oligonucleotide is added and incubation is continued. The 3′OH end of the 'attacking' oligonucleotide can then attack the 5′P-tyrosine covalent link to resolve it. This strand transfer reaction leads to the formation of a new molecule where the attacking oligonucleotide is covalently joined to the cleaved ss REP substrate (Figure 2A4). Pilot experiment with attacking oligonucleotides carrying variable 3′ extremities has shown that the 3′ base is obligatory a C, whereas upstream sequence is less important (not shown). To characterize joint products, purified DNA was used as template for PCR amplification using the attacking oligonucleotide and a primer specific for the REP substrate (Figure 2A5). Typical profile obtained with a ss phagemid substrate carry-

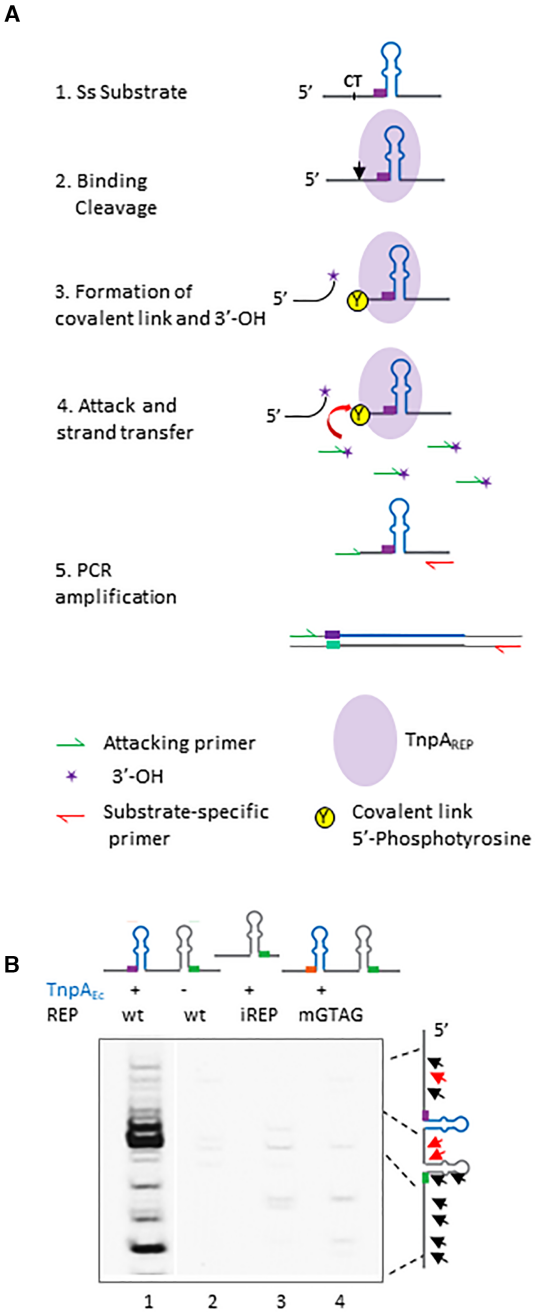

Figure 2. Cleavage Strand Transfer assay (CST). (**A**) Ss DNA substrates (1) were first incubated together with $TnpA_{REP}$ in a reaction buffer leading to their binding and cleavage (2), resulting in the formation of a covalent complex $TnpA_{REP}$ Tyr-5′P and a 3′-OH group (3). Afterwards, an attacking oligonucleotide was added in excess (4), resolving the covalent link and fusing it to the 5′ of cleaved substrate (5). Cleavage sites were mapped by PCR amplification with attacking and substrate-specific primers. Purple oval represents $TnpA_{REP}$, CT/black arrow—cleavage site, purple star - 3′-OH and Y circled in yellow - covalent link Tyr-5′P, respectively. Attacking and substrate specific primers are represented as green and red arrows, respectively. Curved red arrow represents attack by 3′-OH group present on the attacking primer. (**B**) Profile of cleavage sites on ss circular DNA phagemid substrates. The same conditions were used for all the substrates. '–' or '+' indicate no $TnpA_{Ec}$ (lane 2) or with $TnpA_{Ec}$, reactions performed on substrates carrying wild-type REP_{Ec} on a BIME, only iREP or a BIME carrying mutant GTAG (lanes 1–2, 3 and 4 respectively). Black and red arrows (right) represent mapped cleavage sites 5′ and 3′ to REP structure and major cleavage sites in wild-type substrate, respectively.

ing a wild-type REP/BIME is shown in Figure 2B, lane 1, compared to that obtained in the absence of $TnpA_{Ec}$ (lane 2). No significant amplification products were observed using substrates carrying only an iREP or mutations in the essential GTAG motif (Figure 2B, lanes 3 and 4 respectively). In all cases, amplification was specific to wild-type REP/BIME substrate and wild-type $TnpA_{Ec}$, in contrast to catalytic mutant $TnpA_{Ec}$ Y115F (not shown).

The assay was further validated by sequencing the amplification products. As was the case for experiments documented previously, cleavage occurred mainly in proximity, 5′ or 3′of the REP structure (Figure 2B, Supplementary Figure S3A). We also observed discrete distant cleavage sites. In addition, the attacking oligonucleotide was systematically abutted to the T of the C|T cleavage sites confirming that the amplification products were all issued from cleavage and strand joining events (not shown).

CST-based SELEX

To get insight directly into REP structural features potentially important for $TnpA_{REP}$ activity, we took advantage of the CST assay to develop a CST-based SELEX (Systematic Evolution of Ligands by Exponential Enrichment) (25). In contrast to the CST assay described above where phagemid-derived circular ssDNA molecules were used generating multiple cleavage sites 5′ and 3′ to the REP, SELEX substrates are simple oligonucleotides carrying a unique 5′ cleavage site and degenerate zones in the REP defining features (the GTAG motif and the palindrome: bulge, loop). These were incubated with cognate $TnpA_{REP}$ as in the CST assay (Supplementary Figure S3B, R_0). After the first PCR amplification, bulk amplified products were sequenced with a common substrate-specific primer (first round, Supplementary Figure S3B, R_1). For the next round, ss substrates were prepared by asymmetric PCR using an excess of attacking oligonucleotide as described in Materials & Methods (Supplementary Figure S3B, R2). In each round, different 'attacking' oligonucleotides were used, all carrying a 3′C permitting reconstitution of the cleavage site for the next round. Finally, from sequencing data, enrichment of different bases at a given position were estimated by Enrichment factor $E_{N,0}$, calculated as ratio of fractions of a given base at round R_N to that at round R_0: $E_{N,0} = F_N/F_0$. Level of selection (S for score) at each position was then estimated as the variance of $E_{N,0}$ of all bases: $S = V(E_{N,0})$. The calculation method is detailed in Supplementary Materials and an example of this analysis is illustrated in Supplementary Figure S3C.

We first tested the CST-based SELEX to re-examine the importance of the conserved GTAG in the REP_{Ec}. Supplementary Figure S3C shows sequencing profiles obtained with initial substrate (R_0) carrying degenerate bases at the GTAG motif and those obtained at the first round (R_1). Remarkably, the four positions in GTAG motif were selected with high scores at the first round, as illustrated in Figure 3A and Supplementary Figure S3C. This confirmed the crucial role of the motif previously observed: no mutations were tolerated, any substitution abolished binding and cleavage (24, Supplementary Figure S4C lanes 16–18, 19–21 and not shown) and these results therefore validated the test.

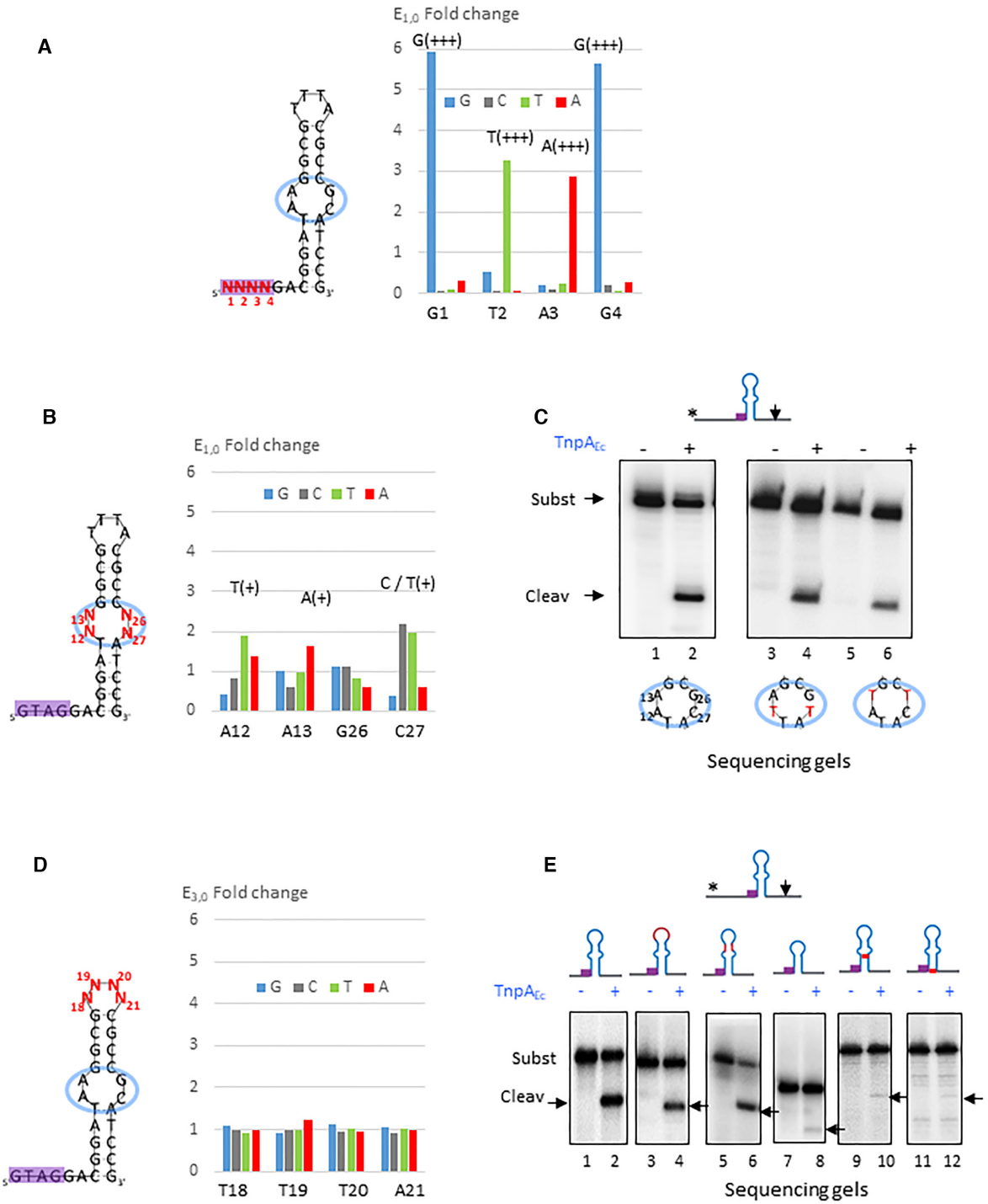

Figure 3. Group 2 *Escherichia coli* REPtron Ec. (**A**) CST-based SELEX and enrichment of the REP_{Ec} GTAG motif. Left: REP_{Ec} structure carrying a 5′ cleavage site (not shown) and degenerate sequence $N_1N_2N_3N_4$ (in red) at the GTAG motif. Right: plot representing Enrichment factor $E_{1,0}$ at the first round R1 of the motif at each position with corresponding scores where G, C, T, A are in blue, grey, green and red respectively. Underneath: initial sequence of the motif. (**B**) Importance of the REP_{Ec} bulged region. Left: REP_{Ec} structure carrying degenerate sequence $N_{12}N_{13}$-$N_{26}N_{27}$ (in red) at the bulged region. Right: plot representing Enrichment factor $E_{1,0}$ of the motif at each position with corresponding scores where G, C, T, A are in blue, grey, green and red, respectively. Below: initial sequence of the motif. (**C**) Cleavage reaction realized with $TnpA_{Ec}$ on wild-type substrate ($A_{12}A_{13}$-$G_{26}C_{27}$) carrying a 3′ cleavage site (lanes 1–2), substrates carrying $\underline{T}_{12}A_{13}$-$G_{26}\underline{T}_{27}$ (lanes 3–4) and $A_{12}\underline{T}_{13}$-$\underline{T}_{26}C_{27}$ (lanes 5–6) respectively. Below: bulged regions are circled in blue where mutated bases are presented in red and cartoons represent corresponding REP structures. (**D**) SELEX of y REP loop sequence. Left: y REP_{Ec} structure carrying degenerate sequence $N_{18}N_{19}N_{20}N_{21}$ (in red) at the loop. Right: plot representing $E_{3,0}$ at the third round R3 of y REP loop at each position where G, C, T, A are in blue, grey, green and red, respectively. Underneath: initial sequence of the y REP loop. (**E**). Cleavage experiment performed with increasing $TnpA_{Ec}$ on wild-type y REP substrate (lanes 1–2), substrate carrying complement of the loop sequence (lanes 3–4), complement of the superior stem (lanes 5–6), substrate deleted for superior half (lanes 7–8), substrates carrying mutations in the conserved positions $T_{11}A$ (lanes 9–10) and $G_{32}C$ (lanes 11–12), respectively.

What has been learned from CST-SELEX on the *E. coli* REPtron Ec

We have shown previously that $TnpA_{Ec}$ is active on the three REP_{Ec} y, z1 and z2 (23). In this section, except indicated otherwise, we generally used oligonucleotides substrates carrying derivatives of the y REP_{Ec}, the most studied at biochemical and structural levels. REP coordinates were kept as used previously (24).

The REP_{Ec} bulge. The conserved mismatches $A_{12}A_{13}$-$G_{26}C_{27}$ are located in the middle of the y REP stem and the C_{27} base is specifically contacted by $TnpA_{Ec}$ (Figure 6B bottom; (24)). Mutations $A_{12}A_{13}$-$T_{26}T_{27}$ or $G_{12}C_{13}$-$G_{26}C_{27}$ introduced to correct the mismatches severely affected activity (24,23). One could therefore expect a significant or exclusive selection of these bases in the CST-based SELEX assay. Instead, while some selections occurred for the three positions $A_{12}A_{13}$ and C_{27}, the enrichments were far from those observed with the GTAG motif (Figure 3B). In particular, both C and T were only moderately enriched at the C_{27} position which is in contact with the protein. The same was observed with conserved positions $A_{12}A_{13}$ where T_{12} and A_{13} were merely enriched with medium scores, respectively (Figure 3B). Medium and low scores could result from the poor selection of independent bases at each position. Alternatively, multiple specific combinations of nucleotides may have been selected. However, bulk Sanger sequencing cannot capture associations between positions and only provide an average picture of the selection process. Since individual selected molecules were not sequenced the analysis cannot inform us directly about synergism or antagonism between substitutions.

To investigate the impact of this 'low selection', we tested substrates carrying substitutions $A_{12}T$, $C_{27}T$, replacing natural mismatch positions by those suggested by SELEX or by other bases $A_{13}T$, $G_{26}T$, both keeping bases unpaired. Cleavage of these variants was maintained as judged by the presence of cleavage products (Figure 3C, compare lanes 3–4, 5–6 to 1–2). Thus, the unpaired state (mispairing in this case) instead of the sequence, seems to be crucial for recognition of the REP_{Ec} by $TnpA_{Ec}$.

The REP_{Ec} stem-loop. Beyond the conserved bulge, hundreds of y, z1 and z2 REP_{Ec}s share several common features (see consensus alignment in Supplementary Figure S2B) including a position in upper stem and several conserved positions in the lower stem, and in particular T_{11} and G_{32} contacted by $TnpA_{Ec}$ (24) (Figure 6B bottom). Among these 3 types of REPs, stem lengths and loop sequences are variable while relatively conserved in each group. To get access to the role of respective loops, we performed CST-SELEX on the 3 types of REPs. No specific enrichment of degenerate loop nucleotides occurred even after several rounds (with $E_{3,0}$ around 1 and low scores for all positions) (Figure 3D, result shown for y and Supplementary Figure S4A-B for z1 and z2 REP_{Ec}, respectively). We further tested the importance of the upper stem sequence and length by mutations. Binding and cleavage of a P^{32}-labelled oligonucleotide substrate for which the loop sequences or the upper stem were swapped to their complement, were still observed, as shown by EMSA (Supplementary Figure S4C, compare lanes 1–3, lanes 4–6 and not shown) and sequencing gel (Figure 3E, lanes 1–2, 3–4 and 5–6), respectively. Similarly, no notable effect was observed upon modification of y REP lower or upper stem to simulate the z1 and z2 structures (not shown). Nevertheless, ablation of the upper stem and loop abolished binding and severely affected cleavage as judged by the absence of retarded complex (Supplementary Figure S4C, lanes 7–9) and reduction of cleavage products (Figure 3E, lanes 7–8). These results suggest a non-specific structural role of the REP_{Ec} upper stem-loop. This is in contrast to the role of the conserved positions T_{11} and G_{32} in the lower stem, which mutations $T_{11}A$ or $G_{32}C$ seriously affected binding (Supplementary Figure S4C, lanes 10–12 and 13–15, respectively) and cleavage (Figure 3E, lanes 9–10 and 11–12, respectively).

Cross-activity. Taken together, these results suggest a relative flexibility in the substrates of $TnpA_{Ec}$. This implies that other REP structures, harboring only few conserved features with REP_{Ec} could be recognized and processed by $TnpA_{Ec}$. Examination of REP structures in two group 2 REPtrons has pointed out some potential common features in REP_{Ec} and REP_{Ma} (Supplementary Figure S5A). Consistently, $TnpA_{Ec}$ exhibits robust activity on REP_{Ma1} and REP_{Ma2} substrates (Supplementary Figure S5B, lanes 1–2 and 3–4, respectively). The importance of the bulge for activity could be confirmed by experiment where mutations introduced to form perfect stem affect $TnpA_{Ec}$ cleavage activity on REP_{Ma1} and REP_{Ma2} substrates (Supplementary Figure S5C, lanes 3–4 compared to 1–2 and lanes 7–8 compared to 5–6).

On top of the crucial GTAG motif, a handful of REP additional structural features appears sufficient to be recognized and processed by $TnpA_{Ec}$.

S. maltophilia REPtron Sm: different strategy to recognize cognate REP

The REPtron Sm includes REPs of 23 nts (REP_{Sm}) composed of an 8-bp perfect palindrome and a 3-nt loop (Figure 1B). Purified Sm $TnpA_{REP}$ ($TnpA_{Sm}$) cleaves REP_{Sm} substrate (an oligonucleotide carrying REP structure and a 3′ cleavage site) at a CT dinucleotide, as shown in Figure 4A (lanes 1–3). No cleavage product was observed with the catalytic mutant derivative $TnpA_{Sm}$ Y130F (lanes 4–6) nor in the presence of a substrate carrying the mutant cleavage site CT-TT (lanes 7–9). As observed for Ec REPtron, the $iREP_{Sm}$ displayed no binding and cleavage activity (not shown). To examine the importance of the conserved GTAG motif and the palindrome features of ss REP_{Sm}, we assayed different ss REP_{Sm} substrates for binding, cleavage *in vitro* and SELEX.

The GTAG motif. $TnpA_{Sm}$ formed specific retarded complex with ss REP_{Sm}, as shown in EMSA experiments (Figure 4B, lanes 1–2). Single mutations in the GTAG motif did not affect the binding profile (Figure 4B, lanes 3–10), showing that, in contrast to TnpAEc (Supplementary Figure S4C, lanes 16–18 and 19–21 and (24)), TnpASm binding to its substrate tolerates mutations in the conserved tetranucleotide. However, these mutations seriously affected cleavage as shown in Figure 4C. Activity was reduced with CTAG mutant and barely detected with GCAG

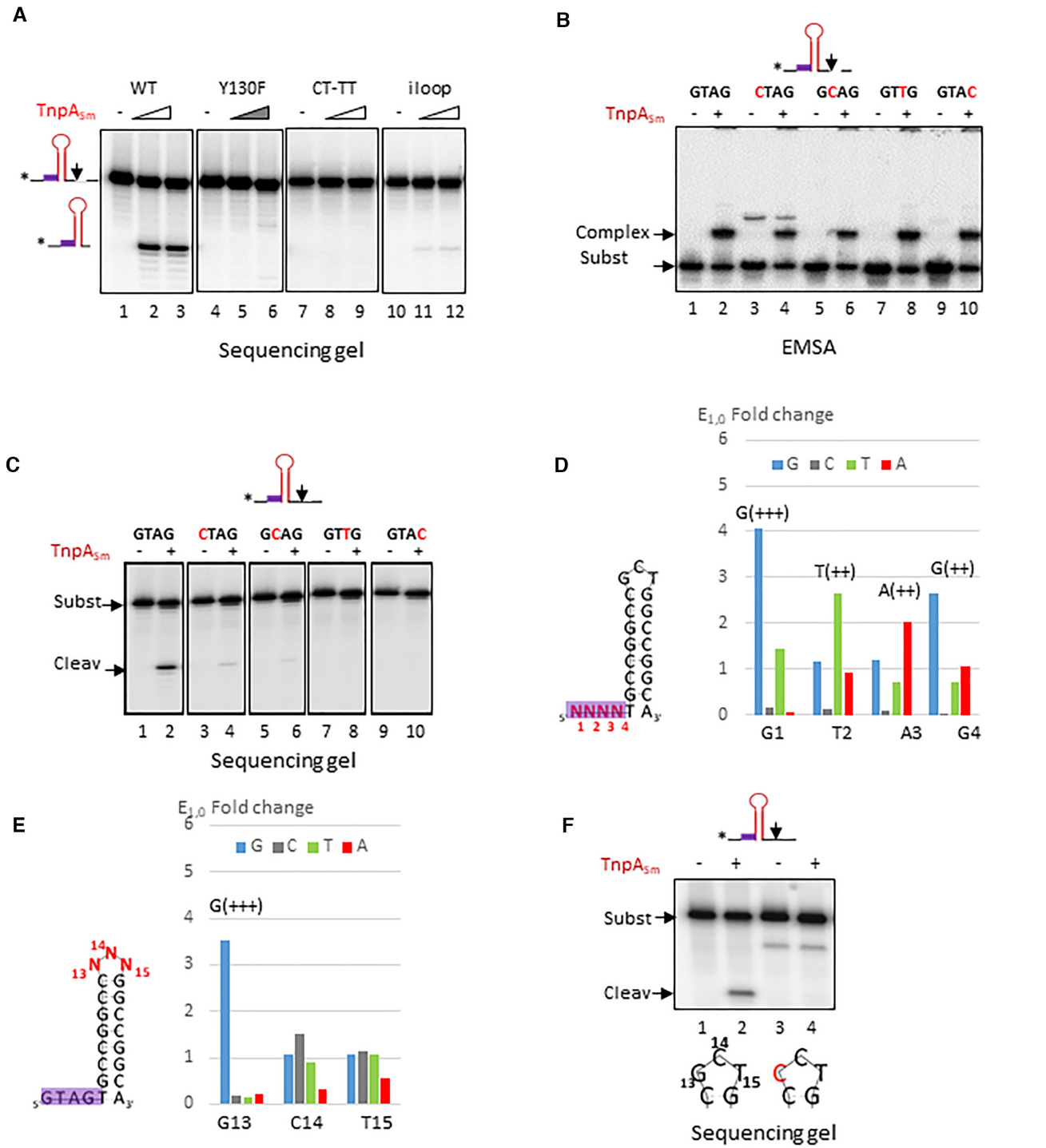

Figure 4. Group 3 *Stenotrophomonas maltophilia* REPtron Sm. (**A**) Cleavage of 55 nts REP_{Sm} substrates performed with increasing concentrations (1 and 4 μM) of wild-type $TnpA_{Sm}$ (lanes 1–3), catalytic mutant derivative $TnpA_{Sm}$ Y130F (lanes 4–6), wild-type $TnpA_{Sm}$ on substrates carrying mutated cleavage site CT-TT and reverse complement of the loop sequence (lanes 7–9 and 10–12, respectively). (**B**) Importance of the GTAG motif for binding. EMSA experiment performed with 2 μM wild-type $TnpA_{Sm}$ on wild-type GTAG substrate (lanes 1–2), CTAG (lanes 3–4), GCAG (lanes 5–6), lanes GTTG (lanes 7–8) and GTAC substrates (lanes 9–10). Mutated positions are indicated in red. (**C**) Importance of the GTAG motif for cleavage. Cleavage experiment performed on a substrate carrying a 3′ cleavage site with 2 μM of wild-type $TnpA_{Sm}$ on wild-type GTAG substrate (lanes 1–2), CTAG (lanes 3–4), GCAG (lanes 5–6), lanes GTTG (lanes 7–8) and GTAC substrates (lanes 9–10). Mutated positions are indicated in red. (**D**) Enrichment of the REP_{Sm} GTAG motif. Left: the same schema as described previously where the $REP_{Sm,}$ structure carrying a 5′ CT cleavage site and a degenerate sequence $N_1N_2N_3N_4$ at the GTAG motif. Right: plot representing $E_{1,0}$ (Enrichment factor) of the motif at each position with corresponding scores. Below: initial sequence of the motif. (**E**) REP_{Sm} loop SELEX and enrichment at the first round. Left: REP_{Sm} structure carrying a degenerate sequence $N_{13}N_{14}N_{15}$ (in red) at the loop. Right: plot representing Enrichment factor $E_{1,0}$ at the first round R1 of the loop at each position where G, C, T, A are in blue, grey, green and red, respectively. High score is indicated for G_{13}. Below: initial sequence of the REP_{Sm} loop. (**F**) Effect of loop mutations on activity. Cleavage of 55 nts REP_{Sm} substrate carrying wild-type $G_{13}C_{14}T_{15}$ or $\underline{C}_{13}C_{14}T_{15}$ loop sequence performed with wild-type $TnpA_{Sm}$ (lanes 1–2 and 3–4, respectively).

substrate (Figure 4C lanes 3–4 and 5–6, respectively compared to wild-type GTAG, lanes 1–2). Mutations in the third and fourth positions completely abolished cleavage (GTTG, lanes 7–8 and GTAC, lanes 9–10). In agreement with these results, in a SELEX experiment, the GTAG motif was selected was selected mainly with good scores (Figure 4D).

The REP_{Sm} stem-loop. In a first series of experiments, we used a mutant carrying a reverse complement of the loop sequence ($G_{13}C_{14}T_{15}$- $A_{13}G_{14}C_{15}$). Cleavage was severely affected, as shown in Figure 4A (lanes 10–12). These mutations also largely compromised binding since no retarded complex was observed by EMSA experiment (not shown), suggesting its critical role in REP recognition. We further investigated the importance of the loop by CST-based SELEX (Figure 4E). Among the 3 bases $G_{13}C_{14}T_{15}$, the G_{13} was largely enriched with high score whereas C_{14} and T_{15} in particular, were not. Accordingly, mutation of a guanine base to a cytosine $G_{13}C$ ($C_{13}C_{14}T_{15}$) abolished cleavage (Figure 4F, compare lanes 1–2 and 3–4), confirming the SELEX result and highlighting the crucial role of this specific position in the REP_{Sm} loop.

To get access to the importance of the REP_{Sm} stem, we introduced mutations mostly by changing nucleotides to their complements by blocs, and subsequently at individual positions. These experiments showed a certain role of the central and upper parts of the stem on cleavage (Supplementary Figure S6A, compare lanes 1–2 with lanes 3–4, 5–6, 9–10, 11–12 and 13–14) although the effect was not drastic. Interestingly, such mutations in three bottom positions improved the cleavage (lanes 7–8). We also tested importance of being a perfect stem by introduction of a mismatch near the middle of the REP_{Sm} stem. These mutations affected or almost eliminated cleavage (Supplementary Figure S6B, compare lanes 4–6, 7–9 to lanes 1–3).

Marinomonas sp. MWYL1 REPtron Mb: 'flexibility' in cleavage site selection

The *Marinomonas* group 3 REPtron Mb comprises small 5 bps perfect palindromic REPs, separated by 2 bases from the GTAG tetranucleotide (Figure 1B). Since $TnpA_{Mb}$ binding to its REP substrate cannot be visualized by EMSA probably due to instability of complexes, here we examined only cleavage activity. Interestingly, the system turned out being more flexible: $TnpA_{Mb}$ (Mb $TnpA_{REP}$) cleaved cognate REP_{Mb} at two sites, CT and CA. A ss DNA substrate of 55 nts carrying these cleavage sites both 5′ and 3′ to the stem-loop exhibited 4 cleavage products (Figure 5A, lanes 1–3). Cleavage sites were confirmed by CST assay and mutational analysis. As expected, no cleavage product was observed with the catalytic mutant derivative $TnpA_{Mb}$ Y125F (lanes 4–6). A substrate carrying the mutant cleavage site CT-TT gave rise to cleavage products at the CA sites only (lanes 7–9).

The GTAG motif: SELEX and role in cleavage site selection. In the case of REPtron Ec, the GTAG tetranucleotide is not only involved in $TnpA_{Ec}$ recognition of REP_{Ec} but also supposed to participate in cleavage site selection (24). Hence we examine the importance of the GTAG motif by SELEX in oligonucleotide substrates carrying 5′ CT or CA cleavage sites separately. We first observed that the selected profile with a CT-carrying SELEX substrate contrasted with the result obtained with REPtron Ec: only the last two positions were strongly enriched with high scores after a single round of enrichment (Figure 5B, left). The profile obtained with CA-carrying substrate showed mainly moderate, more homogenous selection with relatively good scores for the motif (Figure 5B, right).

Cleavage site of IS*200*/IS*605* family members is selected by particular DNA-DNA linear and cross complementarity with guide sequences, tetranucleotide 5′ to the palindromes at left and right IS ends (19) (Supplementary Figure S1). A simple model of REP cleavage site selection would thus involve the GTAG tetranucleotide as a guide sequence (24). Accordingly, CT and CA can be chosen by cross complementarity with A_3G_4 and T_2G_4 respectively (Figure 5C, top). To test this hypothesis, we designed simple 38 nts substrates carrying mutated GTAG variants and a unique cleavage site located 3′ to the stem-loop. The wild-type GTAG substrate was cleaved at CA and CT sites (Figure 5C lanes 1–2 and not shown). Although less efficiently, a substrate carrying a mutation of the third base (GTAG-GTGG) was again cleaved at CA, as expected (lanes 3–4). Changing of GTAG to GTGG resulted in cleavage at CC (lanes 5–6) and to GTAC in cleavage at GA (lanes 7–8) and GT sites (lanes 9–10), respectively. Importantly, no cleavage was detected in absence of the corresponding cleavage site (lanes 11–12).

Thus, different positions of the GTAG motif were selected in substrates carrying CT or CA cleavage sites and although efficacy varied, changing a subset of the motif could modify REP_{Mb} cleavage sites in a predictable way according to two presumed schemas and examples shown in Figure 5C. This confirmed the active role of the motif in cleavage sites selection of this 'flexible' REPtron, in a manner similar to that described for the IS*200*/IS*605* elements (19,26).

The Mb stem–loop. The swap of the entire REP_{Mb} stem to its complement moderately affected cleavage (not shown) indicating a slight role in the REP recognition/activity. Similarly, we further analyzed the importance of different stem portions by the same procedure. We observed a diminution of cleavage activity for mutations of the fourth position in the REP_{Mb} stem, but no effect for mutations of the second and third positions (Figure 5D, compare lanes 1–2 and 7–8 and not shown). Similarly to the REPtron Sm, introduction of a mismatch in the REP_{Mb} stem seriously diminished cleavage activity (Figure 5E, compare lanes 3–4 and 5–6 to lanes 1–2). We then examined the importance of the REP_{Mb} loop by SELEX. Experiments were performed separately on CT- or CA- carrying substrates with a degenerate loop. For both substrates, three among 4 positions (T_{12}, T_{13} and A_{15}) were strongly enriched with good and excellent scores (Figure 5F and not shown). Accordingly, negative values of $\log2(E_{1,0})$ (for $E_{1,0}$ below 1), clearly illustrated exclusion of the rest in these three positions T_{12}, T_{13} and A_{15} (Supplementary Figure S7). These counter-selections were otherwise confirmed by mutational analysis shown in Figure 5G. Mutations in the first and second po-

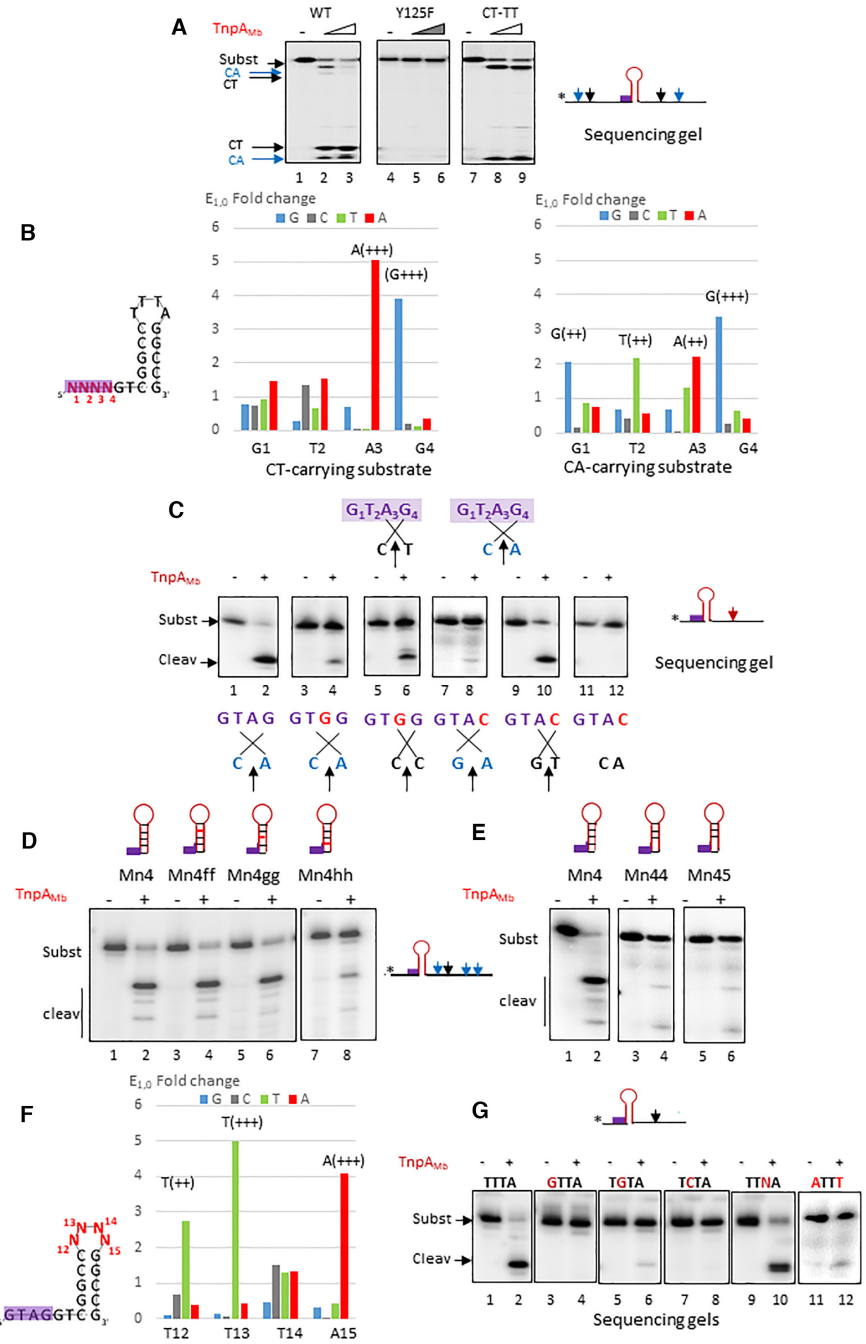

Figure 5. Group 3 *Marinomonas* sp. MWLY1 REPtron Mb. (**A**) $TnpA_{Mb}$ cleaves cognate REP at CT and CA. Cleavage experiment performed with increasing concentrations of wild-type $TnpA_{Mb}$ on 55 nts wild-type substrate (lanes 1–3), $TnpA_{Mb}$ catalytic mutant derivative Y125F on wild-type substrate (lanes 4–6) and substrate carrying mutations CT-TT at two CT sites (lanes 7–9). CA and CT cleavage products are shown by blue and black arrows, respectively. (**B**) REP_{Mb} GTAG SELEX on CT-carrying substrate (left) and on CA-carrying substrate (right). The same schema as described previously is shown, REP_{Mb} carrying a 5′ CT or CA cleavage site and degenerate sequence $N_1N_2N_3N_4$ (in red) at the GTAG motif where G, C, T, A are in blue, grey, green and red, respectively. Scores are indicated at corresponding positions. Underneath: initial sequence of the motif. (**C**) Model of CT or CA-cleavage sites selection based on IS*608* model (top). GTAG mutations and cleavage sites selection (bottom). Cleavage of 35 nts simple substrates carrying mutations in the GTAG motif and a 3′ unique cleavage sites: wild-type GTAG and CA (lanes 1–2), GTGG and CA (lanes 3–4), GTGG and CC (lanes 5–6), GTAC and GA (lanes 7–8), GTAC and GT (lanes 9–10), GTAC and CA (lanes 11–12). Schemas of cleavage sites potential selection are shown below each gel. (**D**) Role of REP_{Mb} stem. Cleavage of 39 nts REP_{Mb} wild-type substrate by $TnpA_{Mb}$ (lanes 1–2), substrates with mutated 2nd (lanes 3–4), 3rd (lanes 5–6) and 4th positions (lanes 7–8) respectively. Top: Cartoons representing REP structures with mutated positions in red. (**E**) Role of REP_{Mb} perfect stem. Cleavage of 39 nts REP_{Mb} wild-type substrate by $TnpA_{Mb}$ (lanes 1–2), substrates with mismatch G-A (lanes 3–4) or T-C (lanes 5–6) at 4th position, respectively. Top: Cartoons representing wild-type and REP structures with mutated positions. (**F**). REP_{Mb} loop sequence at R1 selection ($E_{1,0}$) with corresponding scores where G, C, T, A are in blue, grey, green and red, respectively. The same schema as described previously with $N_{12}N_{13}N_{14}N_{15}$ degenerate loop sequence. Underneath: initial sequence of the motif. (**G**) Effect of REP_{Mb} loop mutations on activity. Cleavage experiments on 39 nts simple CT-carrying substrate with wild-type loop $T_{12}T_{13}T_{14}A_{15}$ (lanes 1–2), $\underline{G}_{12}T_{13}T_{14}A_{15}$ (lanes 3–4), $T_{12}\underline{G}_{13}T_{14}A_{15}$ (lanes 5–6), $T_{12}\underline{C}_{13}T_{14}A_{15}$ (lanes 7–8), $T_{12}T_{13}\underline{N}_{14}A_{15}$ (lanes 9–10) and $\underline{A}_{12}T_1T_{14}\underline{T}_{15}$ (lanes 11–12), respectively.

sitions ($\underline{G}_{12}T_{13}T_{14}A_{15}$ and $T_{12}\underline{G}_{13}T_{14}A_{15}$) greatly reduced cleavage (Figure 5G, lanes 3–4 and lanes 5–6, compared to lanes 1–2). Also, the replacement $T_{13}C$ ($T_{12}\underline{C}_{13}T_{14}A_{15}$) completely abolished activity (lanes 7–8) while substrate carrying the 14th base degenerate ($T_{12}T_{13}\underline{N}_{14}A_{15}$) exhibited wild-type behavior (lanes 9–10). Finally, exchange of T_{12} and A_{15} ($\underline{A}_{12}T_{13}T_{14}\underline{T}_{15}$) or individual substitutions $T_{12}A$ ($\underline{A}_{12}T_{13}T_{14}\underline{A}_{15}$) or $A_{15}T$ ($T_{12}T_{13}T_{14}\underline{A}_{15}$) also compromised activity, as shown in Figure 5G, lanes 11–12 and not shown.

These results confirmed the crucial role of three positions in the REP_{Mb} loop in cleavage activity and suggest that two bases T_{12} A_{15} are complementary in the REP_{Mb} structure and might be considered as part of the stem.

DISCUSSION

Our analysis demonstrated that $TnpA_{REP}$ of the two groups employ diverse strategies to recognize their REP substrates. Clearly both REP components, the GTAG tetranucleotide motif and the palindrome, were involved in $TnpA_{REP}$ activity but their respective impacts varied in each system. In Figure 6A, we summarize the importance of these features. While GTAG is instrumental in REPtron Ec, mutations are largely tolerated in REPtron Sm for binding and in REPtron Mb for cleavage (and by deduction for binding). Although involvement of the GTAG motif in cleavage site selection has been suggested for REPtron Ec, its role was not experimentally supported. Interestingly, the REP_{Mb} tetranucleotide motif was differently selected in CA or CT carrying substrates, probably reflecting their distinct contribution to respective cleavage sites selection. The role of loop sequences is also different for representatives of the two groups. No mutations were tolerated in certain REP_{Mb} or REP_{Sm} loop positions, while only a non-specific structural role was suggested for the REP_{Ec} loop.

$TnpA_{REP}$ of the two groups

Catalytic center and C-term tail. Groups 2 and 3 REPtrons differ by their encoded $TnpA_{REP}$ and corresponding REPs. As shown by an alignment performed on a limited collection of $TnpA_{REP}$ (Figure 6B), the catalytic center composed of the metal coordination module (HuH motif and other additional residues (24)) and the catalytic Tyr is well conserved in both groups. Some differences are found in the N-term and C-term portions: group 3 members include several supplementary residues in N-term whereas group 2 members carry a C-term extension of about 20 residues, comprising a short helix α5 and an unstructured region in the case of $TnpA_{Ec}$ (Figure 6B). The helix α5 and downstream adjacent region appeared to be important in $TnpA_{Ec}$ activity since derivatives Δ131 and Δ144 (deletions of 34 and 21 C-terminal residues, respectively) exhibit serious defects in binding and cleavage (data not shown). However deletion of 13 extreme C-terminal residues resulted in a mutant, Δ152, with higher activity than the wild-type (24), suggesting a regulatory function for these residues. In the group 3, the C-term part comprises also a short helix of unknown function.

Contacts with REP. The REP_{Ec} GTAG motif, which is exclusively selected in SELEX and which tolerates no substitution for binding and catalytic activity, is heavily contacted by $TnpA_{Ec}$ protein residues (group 2). These residues are distributed in the regions comprising β1 and surrounding β4 and also the C-terminal extremity (24) (Figure 6B). While these residues are well conserved in group 2, only some (Q95, D100 and R104) are relatively conserved in the group 3. In particular, G160 and E161, situated in the $TnpA_{Ec}$ C-term tail and absent in the group 3, contact the last two bases of the GTAG motif. These differences may partly explain the discrepancy in GTAG requirement in the two groups.

In REP_{Ec} (group 2), the conserved mismatches forming a bulge in the middle of the stem $A_{12}A_{13}$-$G_{26}C_{27}$ were also important since mutations recreating perfect palindrome affected activity, the C_{27} is specifically contacted by the residue K82 situated in the conserved DNA binding α3 helix (Figure 6B, (24)). Nevertheless, these positions were not or only moderately selected by SELEX suggesting that different combinations of nucleotides are possible. And in the case of C_{27}, we obtained a mixture C/T suggesting that a pyrimidine might be required at this position. Concerning group 3 REPs, exclusive selection of unique loop positions G_{13} (Sm), and T_{13} (Mb) and impact of mutations on activity demonstrated their crucial role (Figures 4 and 5). Since no structural data are available, we can only speculate relative to contacts with cognate $TnpA_{REP}$. In spite of discrepancy, some parallel might be made between loop positions in small REPs of group 3 and unpaired positions in group 2 REP and residues on the equivalent DNA binding helix α3 might be responsible for these contacts. The same helix and downstream region might mediate cognate $TnpA_{REP}$ binding to the group 3 REP stems as observed for $TnpA_{Ec.}$

Binding to folded ssDNA hairpin

$TnpA_{REP}$, as $TnpA_{IS200/IS605}$, recognize their ss DNA REP substrates in a strand-specific manner. Only REP with characteristic features is bound and processed, iREP is not. In the group 3 REP, the conserved motif GTAG is clearly involved in strand discrimination, while its role in group 2 is more limited. Furthermore, the effect of single stranded features (loop or irregular zone as mismatches, bulge) is undeniable.

These properties echo those displayed by some proteins encoded by mobile genetic elements working on ss folded DNA such as Integrases IntI encoded by Integron, plasmid Relaxases and ss DNA Transposases $TnpA_{IS200/IS605}$. For conjugative transfer, the Relaxase recognizes *oriT* as a single stranded folded hairpin (27). While contacts with stem remain non-specific, it establishes specific contacts with ss DNA cleavage region downstream of the hairpin. In the recombination reaction between the integron *attC* and *attI* sites, the ds DNA site *attC* is a ss folded structure from the bottom strand, reconstituting ds recombination site (28). In the crystal structure, Int establishes specific contacts with several flipped out bases in the *attC* site and these interactions are primordial for recombination (29). Moreover, efficient insertion of integron cassette is also influenced by two other unpaired regions of *attC* recombination sites (30). Recently, the impact of these structural speci-

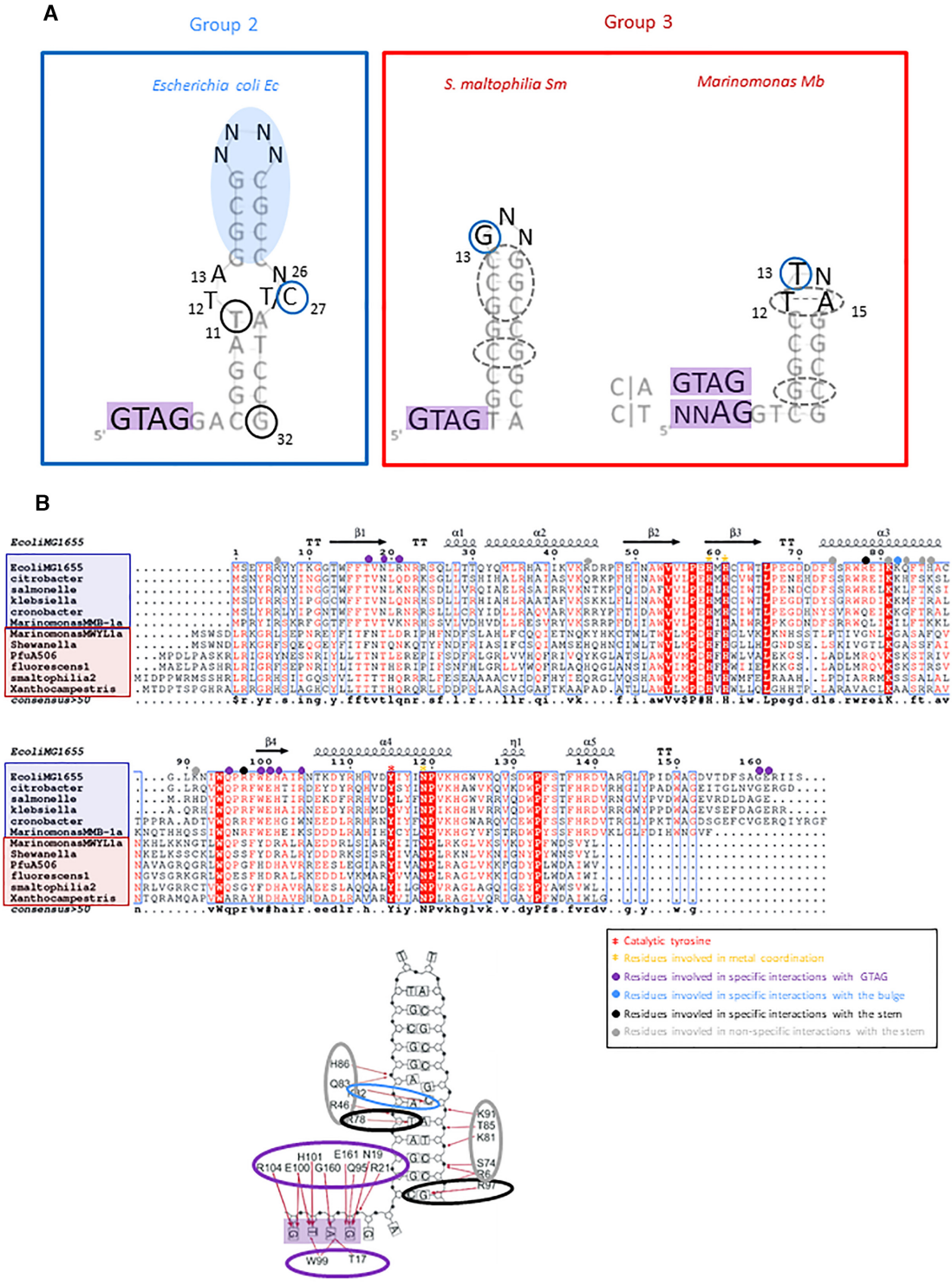

Figure 6. REP structural determinants in three models and putative contacts $TnpA_{REP}$/REP. (**A**) Zones important for $TnpA_{REP}$ activity. Summary of the roles of REP structural components in three models, positions tested by CST-based SELEX are in black, where font size reflects enrichment score. Left: Selected bases in the REP_{Ec} bulge region and non-specific structural upper half (blue oval) is presented. Conserved positions contacted by $TnpA_{Ec}$ C_{27} (bulge) and T_{11}, G_{32} (lower stem) are circled in blue and black, respectively, where the latter importance was confirmed experimentally. Right: stem important zones are circled with dotted lines, loop key positions revealed by SELEX are circled in blue. REP_{Mb} GTAG SELEX results of CA and CT-carrying substrates are boxed separately. REP_{Mb} important positions T_{12} and A_{15} are presented paired and circled with dotted lines, as suggested by mutational analysis. (**B**) Top: Alignment of groups 2 and 3 $TnpA_{REP}$ (boxed in blue and red, respectively) based on $TnpA_{Ec}$ structural data. Catalytic tyrosine and HuH motif are indicated by red and orange-coloured stars, respectively. $TnpA_{Ec}$ residues involved in specific contacts with the GTAG motif, specific interaction with the bulge, specific and non-specific contacts with REP_{Ec} stem are indicated by purple, blue, black and grey points, respectively. Bottom: $TnpA_{Ec}$ residues contacting minimal y REP_{Ec} structure (24) where the same colour code is used: residues contacting specifically GTAG (boxed in purple), bulged C_{27} (in light blue) and stem specific positions T_{11}, G_{32} (in black) and stem non-specific interactions (in grey), respectively.

ficity determinants of integron cassette has been refined using synthetic biology combined with large scale mutagenesis, next-generation sequencing and machine learning (31). This powerful approach will be a valuable tool to reexamine and to get a global view of specificity determinants and synthetic evolution pathways in diverse systems including REPtrons.

$TnpA_{REP}$ are so far the closest relatives of $TnpA_{IS200/IS605}$, among which transposases of IS*608* and IS*Dra2* are the most studied. To recognize the REP correct structure, $TnpA_{REP}$ proteins contact loop or irregularities in the palindrome stem, as do ss transposases. IS*608* unpaired base T_{17} is sandwiched between aromatic residues in a hydrophobic pocket, whereas the T_{10} in the loop is specifically contacted by two residues (17). Similarly, IS*Dra2* transposase displayed contacts with T_{14} in the loop and a mismatched base within the stem (18). Thus, $TnpA_{REP}$ proteins employ alternatively these binding determinants in combination with the conserved tetranucleotide GTAG. The last feature clearly distinguishes $TnpA_{REP}$ from ss transposases that mostly contact exclusively the palindromes.

Cleavage sites selection

Left and right cleavage sites of the IS*200*/IS*605* family members are selected via a network of peculiar complementary interactions with corresponding 'guide' sequences, which are tetranucleotides 5′ to the palindromes (19) (Supplementary Figure S1). Consequently, IS*608* cleavage sites could be modified by changing the corresponding 'guide' sequences, resulting also in retargeting of the IS (26). The position of the GTAG tetranucleotide in REPs could be equivalent to the 'guide' sequences. According to the proposed model of cleavage site selection based on examples of IS*608* and IS*Dra2*, the common CT and the Mb CA cleavage sites would be chosen via interactions with subsets of the conserved GTAG. $TnpA_{Mb}$ turned out to be more flexible and cleaves REP substrate at both CT and CA sites. Thanks to this flexibility, we could explore this question and manage to vary REP_{Mb} CT and CA cleavage sites by changing certain positions in the GTAG motif. Although in these experiments the cleavage sites could be changed by that simple way, cleavage efficiency varies and it is not excluded that other factors would be involved.

In the cases of REPtrons Ec and Sm, similar attempts to change cleavage sites did not succeed (not shown). The REPtron Ec is known not to tolerate any GTAG mutation. In the case of Sm, while GTAG mutants still form complexes with $TnpA_{Sm}$, they severely reduced cleavage, in particular when mutations concern the last two positions, consistent with their postulated role in cleavage site selection. We suppose that the CT site is indeed selected by the GTAG motif but that, in the case of the REPtron Ec, the GTAG is 'protected' from mutation by specific contacts with the protein, as shown by the structure. Alternatively, $TnpA_{Ec}$ or $TnpA_{Sm}$ could also accommodate CT dinucleotide into the catalytic site. This information was missing in the available structure.

REPtrons and potential evolutionary route

REPtrons and IS*200*/IS*605* family members share major features. They exhibit an equivalent genetic structure in which coding sequences are bordered by palindromes, and encode proteins with a similar catalytic domain. Large scale phylogenetic analyses have confirmed the evolutionary relationship between $TnpA_{REP}$ and $TnpA_{IS200/IS605}$ (13,20). $TnpA_{REP}$ have been proposed to originate from ancient $TnpA_{IS200/IS605}$ ancestors in Enterobacteria and Pseudomonas where $tnpA_{REP}$ are the most widespread. Alternatively, this distribution may reflect their successful establishment following arrival via horizontal transfer in these bacterial groups (ISfinder https://isfinder.biotoul.fr/), (32).

Our results here suggest that these two $TnpA_{REP}$ groups co-evolve with their respective REP sequences. On the other hand, this does not seem to be the case with the IS*200*/IS*605* family, which includes two subgroups, one carries palindromes with irregularities (e.g. IS*608* and IS*Dra2*) whereas another one is associated with perfect palindromic ends (e.g. IS*200*, IS*1451*). Yet $TnpA_{IS200/IS605}$ appear very homogenous, no distinction being observable in corresponding transposases sequences (ISfinder). It will be interesting to know whether the two described REPtrons groups here have evolved from a common ancestor, common with IS*200*/IS*605* family members or not.

In spite of the close relationship between $TnpA_{REP}$ and $TnpA_{IS200/IS605}$, while $tnpA_{IS200/IS605}$ exhibit typical behavior of IS transposase genes, $tnpA_{REP}$ are, in many respects, very close to housekeeping genes (13), supporting the previous consideration of $TnpA_{REP}$ as the first described bacterial domesticated transposases. The maintenance of $tnpA_{REP}$ in bacterial genomes also implies that they have been coopted to fulfil functions benefic to the host cell. Diverse documented functions of REP sequences in cell physiology suggest their roles in improving fitness of bacterial host in a given niche or environment. This notion has been reinforced by a recent genome-wide CRISPRi analysis in *E. coli* using the catalytic null mutant of the Cas9 RNA-guided nuclease (CRISPR-dCas9) for silencing genes of interest (33). Interestingly, this study has revealed fitness defect caused by dCas9 binding at different REP sequences. Works are in progress to decipher the dissemination pathway of these important elements.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank M. Chandler, Y. Quentin, G. Fichant and members of the Gedy team for fruitful discussions, F. Cornet, P. Rousseau, M. Campos for critical reading of the manuscript, M. Campos for help in SELEX quantification, C. Hoareau and A. Reynaud for initiating the study.

FUNDING

Centre National de Recherche Scientifique (CNRS, France); Agence National pour la Recherche (ANR)

MOBISING Blanc SVSE8; Fédération de Recherche en Biologie à Toulouse (FRBT). Funding for open access charge: MOBISING Blanc SVSE8.
Conflict of interest statement. None declared.

REFERENCES

1. Delihas,N. (2008) Small mobile sequences in bacteria display diverse structure/function motifs. *Mol. Microbiol.*, **67**, 475–481.
2. Treangen,T.J., Abraham,A.L., Touchon,M. and Rocha,E.P. (2009) Genesis, effects and fates reveals a smooth evolutionary transition during functional innovation ofrepeats in prokaryotic genomes. *FEMS Microbiol. Rewiews*, **33**, 539.
3. Stern,M.J., Ames,G.F., Smith,N.H., Robinson,E.C. and Higgins,C.F. (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell*, **37**, 1015–1026.
4. Higgins,C.F., McLaren,R.S. and Newbury,S.F. (1988) Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. *Gene*, **72**, 3–14.
5. Espeli,O. and Boccard,F. (1997) In vivo cleavage of Escherichia coli BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site. *Mol. Microbiol.*, **26**, 767–777.
6. Espeli,O., Moulin,L. and Boccard,F. (2001) Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *J. Mol. Biol.*, **314**, 375–386.
7. Bachellier,S., Clement,J.M. and Hofnung,M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol.*, **150**, 627–639.
8. Rocco,F., De Gregorio,E. and Di Nocera,P.P. (2010) A giant family of short palindromic sequences in Stenotrophomonas maltophilia. *FEMS Microbiol. Lett.*, **308**, 185–192.
9. Liang,W., Rudd,K.E. and Deutscher,M.P. (2015) A role for REP sequences in regulating translation. *Mol. Cell*, **58**, 431–439.
10. Liang,W. and Deutscher,M.P. (2016) REP sequences: mediators of the environmental stress response? *RNA Biol.*, **13**, 152–156.
11. Nunvar,J., Huckova,T. and Licha,I. (2010) Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. *BMC Genomics*, **11**, 44.
12. Nunvar,J., Licha,I. and Schneider,B. (2013) Evolution of REP diversity: a comparative study. *BMC Genomics*, **14**, 385.
13. Bertels,F., Gallie,J. and Rainey,P.B. (2017) Identification and characterization of domesticated bacterial transposases. *Genome Biol. Evol.*, **9**, 2110–2121.
14. Chandler,M., de la Cruz,F., Dyda,F., Hickman,A.B., Moncalian,G. and Ton-Hoang,B. (2013) Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat. Rev. Microbiol.*, **11**, 525–538.
15. He,S., Corneloup,A., Guynet,C., Lavatine,L., Caumont-Sarcos,A., Siguier,P., Marty,B., Dyda,F., Chandler,M. and Ton Hoang,B. (2015) The IS200/IS605 family and "Peel and Paste" single-strand transposition mechanism. *Microbiol. Spectr.*, **3**, https://doi.org/10.1128/microbiolspec.MDNA3-0039-2014.
16. Guynet,C., Hickman,A.B., Barabas,O., Dyda,F., Chandler,M. and Ton-Hoang,B. (2008) In vitro reconstitution of a single-stranded transposition mechanism of IS608. *Mol. Cell*, **29**, 302–312.
17. Ronning,D.R., Guynet,C., Ton-Hoang,B., Perez,Z.N., Ghirlando,R., Chandler,M. and Dyda,F. (2005) Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol. Cell*, **20**, 143–154.
18. Hickman,A.B., James,J.A., Barabas,O., Pasternak,C., Ton-Hoang,B., Chandler,M., Sommer,S. and Dyda,F. (2010) DNA recognition and the precleavage state during single-stranded DNA transposition in D. radiodurans. *EMBO J.*, **29**, 3840–3852.
19. Barabas,O., Ronning,D.R., Guynet,C., Hickman,A.B., Ton-Hoang,B., Chandler,M. and Dyda,F. (2008) Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. *Cell*, **132**, 208–220.
20. Quentin,Y., Siguier,P., Chandler,M. and Fichant,G. (2018) Single-strand DNA processing: phylogenomics and sequence diversity of a superfamily of potential prokaryotic HuH endonucleases. *BMC Genomics*, **19**, 475.
21. Bertels,F. and Rainey,P.B. (2011) Curiosities of REPINs and RAYTs. *Mob. Genet. Elements*, **1**, 262–268.
22. Bertels,F. and Rainey,P.B. (2011) Within-genome evolution of REPINs: a new family of miniature mobile DNA in bacteria. *PLos Genet.*, **7**, e1002132.
23. Ton-Hoang,B., Siguier,P., Quentin,Y., Onillon,S., Marty,B., Fichant,G. and Chandler,M. (2012) Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. *Nucleic Acids Res.*, **40**, 3596–3609.
24. Messing,S.A., Ton-Hoang,B., Hickman,A.B., McCubbin,A.J., Peaslee,G.F., Ghirlando,R., Chandler,M. and Dyda,F. (2012) The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. *Nucleic Acids Res.*, **40**, 9964–9979.
25. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
26. Guynet,C., Achard,A., Hoang,B.T., Barabas,O., Hickman,A.B., Dyda,F. and Chandler,M. (2009) Resetting the site: redirecting integration of an insertion sequence in a predictable way. *Mol. Cell*, **34**, 612–619.
27. Guasch,A., Lucas,M., Moncalian,G., Cabezas,M., Perez-Luque,R., Gomis-Ruth,F.X., de la Cruz,F. and Coll,M. (2003) Recognition and processing of the origin of transfer DNA by conjugative relaxase TrwC. *Nat. Struct. Biol.*, **10**, 1002–1010.
28. MacDonald,D., Demarre,G., Bouvier,M., Mazel,D. and Gopaul,D.N. (2006) Structural basis for broad DNA-specificity in integron recombination. *Nature*, **440**, 1157–1162.
29. Bouvier,M., Demarre,G. and Mazel,D. (2005) Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J.*, **24**, 4356–4367.
30. Nivina,A., Escudero,J.A., Vit,C., Mazel,D. and Loot,C. (2016) Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of attC recombination sites. *Nucleic Acids Res.*, **44**, 7792–7803.
31. Nivina,A., Grieb,M.S., Loot,C., Bikard,D., Cury,J., Shehata,L., Bernardes,J. and Mazel,D. (2020) Structure-specific DNA recombination sites: design, validation, and machine learning-based refinement. *Sci. Adv.*, **6**, eaay2922.
32. Siguier,P., Filee,J. and Chandler,M. (2006) Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.*, **9**, 526–531.
33. Rousset,F., Cui,L., Siouve,E., Becavin,C., Depardieu,F. and Bikard,D. (2018) Genome-wide CRISPR-dCas9 screens in *E. coli* identify essential genes and phage host factors. *PLos Genet.*, **14**, e1007749.