9964–9979 *Nucleic Acids Research, 2012, Vol. 40, No. 19* *Published online 9 August 2012*
doi:10.1093/nar/gks741

The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease

Simon A. J. Messing[1], Bao Ton-Hoang[2], Alison B. Hickman[1], Andrew J. McCubbin[3], Graham F. Peaslee[3], Rodolfo Ghirlando[1], Michael Chandler[2] and Fred Dyda[1,*]

[1]Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA, [2]Laboratoire de Microbiologie et Génétique Moléculaires Centre National de la Recherche Scientifique, 118 Route de Narbonne, 31062, Toulouse Cedex, France and [3]Chemistry Department, Hope College, 35 E. 12th Street, Holland, MI 49423, USA

Received May 1, 2012; Revised June 20, 2012; Accepted July 11, 2012

ABSTRACT

Extragenic sequences in genomes, such as microRNA and CRISPR, are vital players in the cell. Repetitive extragenic palindromic sequences (REPs) are a class of extragenic sequences, which form nucleotide stem-loop structures. REPs are found in many bacterial species at a high copy number and are important in regulation of certain bacterial functions, such as Integration Host Factor recruitment and mRNA turnover. Although a new clade of putative transposases (RAYTs or $TnpA_{REP}$) is often associated with an increase in these repeats, it is not clear how these proteins might have directed amplification of REPs. We report here the structure to 2.6 Å of $TnpA_{REP}$ from *Escherichia coli* MG1655 bound to a REP. Sequence analysis showed that $TnpA_{REP}$ is highly related to the IS*200*/IS*605* family, but in contrast to IS*200*/IS*605* transposases, $TnpA_{REP}$ is a monomer, is auto-inhibited and is active only in manganese. These features suggest that, relative to IS*200*/IS*605* transposases, it has evolved a different mechanism for the movement of discrete segments of DNA and has been severely down-regulated, perhaps to prevent REPs from sweeping through genomes.

INTRODUCTION

For many years, a large part of the extragenic sequence in genomes was thought to be essentially silent and devoid of function, hence the popular term 'junk DNA'. This paradigm changed considerably over the last two decades, as it has become apparent that these sequences often encode unexpected functions as evidenced by the discovery of microRNAs and their role in gene regulation (1,2), and by the identification of a new class of short palindromic repeats, known as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), which are critical in prokaryotic immunity (3–5).

Repetitive extragenic palindromic sequences (REPs) are a distinct class of abundant repeats important in regulation of certain bacterial functions. REPs are known to interact with several partners, by providing binding sites for proteins such as Integration Host Factor and DNA polymerase I, and providing the necessary cleavage sites for DNA gyrase to unwind DNA (6–8). REPs also increase mRNA stability and can cause transcription termination (9,10). These REP functions are also exhibited to some degree in other extragenic prokaryotic DNA elements, such as the Correia elements in *Neisseria gonorrhoeae*, or RUP elements in *Streptococcus pneumoniae* (11–14). Discovered nearly 30 years ago in enteric bacteria (15,16), REPs are 35–40 nt long stem-loop structures often organized into larger units called bacterial interspersed mosaic elements (BIMEs) (17–19). BIMEs comprise two REPs in inverse orientation separated by a linker sequence (shown in Figure 1a for *Escherichia coli* MG1655). One is called REP and the second inverted sequence is designated an iREP. REPs are found dispersed throughout the chromosome in many bacterial species, often in high copy number. They represent, for example, up to 1% of *E. coli* chromosomes (20–24). They are so frequent that their presence serves as the basis for 'REP-PCR', a method for the rapid identification of bacterial strains (25).

Hints as to how REPs have come to populate some bacterial genomes with such high frequency were

*To whom correspondence should be addressed. Tel: +301 402 4496; Fax: +301 496 0201; Email: fred.dyda@nih.gov

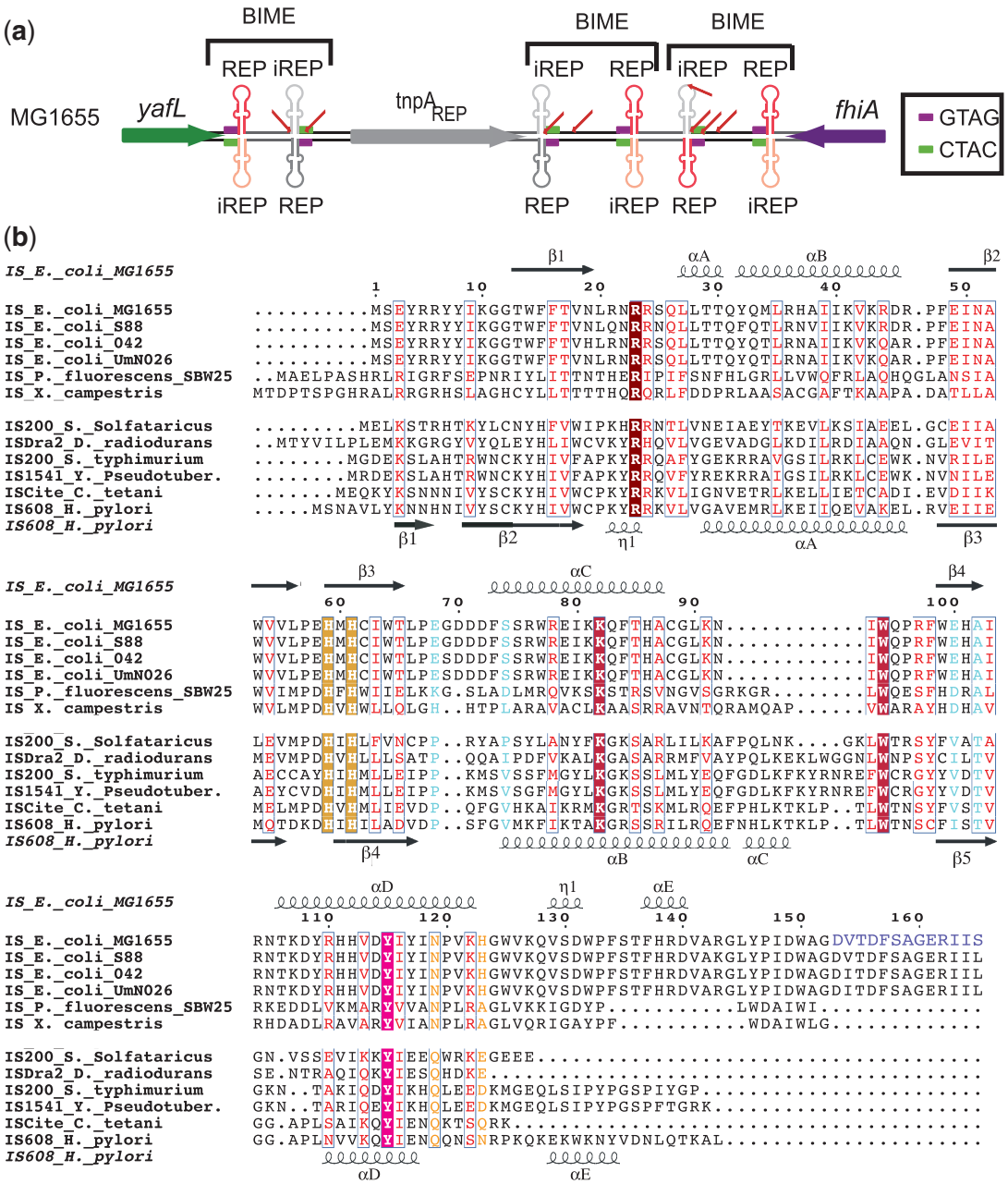

Figure 1. REP organization and IS*200*/IS*605* family alignment. (**a**) Schematic representation of the *E. coli* MG1655 REPtron displaying the organization of BIMEs and their respective REPs and iREPs. The y REP is in red, the y iREP in orange, z1 REP in blue and the z1 iREP in light blue. The y and z1 nomenclature preceding REP or iREP define the consensus nucleotide sequence of the hairpin (27). The 5′ GTAG tetranucleotide is represented by a purple box at the foot of the REP, and the CT dinucleotide cleavage sites previously established in reference 27 are marked by red arrows. (**b**) Multiple sequence alignment of prominent members of the IS*200*/IS*605* family with key members of the new clade. The histidines of the HUH motif are highlighted in orange, and the catalytic tyrosine in magenta. Other residues coordinating the divalent cation in the new clade are in orange script, residues involved in $TnpA_{IS608}$ dimerization in cyan, C-terminal extension of $TnpA_{REP}$ in purple and residues sharing identity throughout are highlighted in red boxes.

obtained from analyses of the genetic regions around REPs, which revealed a new clade of putative transposases, termed REP-associated tyrosine transposases (RAYTs) or $TnpA_{REP}$ (26,27). RAYTs are found in a variety of species at a species-specific single locus, always flanked by BIMEs. Phylogenetic analysis of *E. coli* and *Shigella* indicated RAYTs were acquired early in species radiation (27). In principle, as transposases can cleave and recombine DNA segments, this could explain the spread of REPs in their respective genomes. Consistent with this hypothesis, RAYT presence is correlated with a general increase in REP copy number (26). In addition, recent *in vitro* studies on the $TnpA_{REP}$ from *E. coli* MG1655 showed that it can cleave specific DNA segments if a REP is present and is also able to recombine BIME fragments (27). Furthermore, evidence that REPs may be

capable of excision was obtained from *Pseudomonas fluorescens* (28).

Sequence analysis of the RAYT clade reveals that its members are related to the Y1 transposases of the IS*200*/IS*605* family of ssDNA transposases. The IS*200*/IS*605* family transposases are members of the HUH superfamily of ssDNA nucleases characterized by a highly conserved His-hydrophobic-His (HUH) motif and a single catalytic tyrosine (29). Members of the HUH superfamily are involved in a wide variety of DNA transactions that use ssDNA substrates, such as replication initiation in certain ssDNA viruses, plasmid conjugation, initiation of rolling circle replication and DNA transposition (29–33). The HUH motif (34) is responsible for providing two of the ligands coordinating a single divalent metal ion cofactor, which binds and polarizes the scissile phosphate group of the DNA, aligning it correctly for nucleophilic attack by the catalytic tyrosine.

Biochemical and structural data from a Y1 transposase of the IS*200*/IS*605* family, encoded by the insertion sequence (IS) IS*608* from *Helicobacter pylori*, $TnpA_{IS608}$, permitted a detailed description of the IS*608* transposition cycle (35–37). ISs are the simplest form of mobile DNA that undergo transposition (38). Key to the mobility of the IS*200*/IS*605* family is the ability of the transposase to bind, cleave and rejoin (or recombine) specific ssDNA segments. Y1 transposases recognize and bind hairpins formed by imperfect palindromes (IP) located very close to the ends of the IS. Cleavage of the bound DNA strand at each IS end is mediated by an active-site tyrosine residue, and results in formation of a covalent 5′ phosphotyrosine linked intermediate on one side of the ssDNA break and a free 3′-OH group on the other. One of the most unusual features of mobile elements from this family is that site-specific cleavages at the IS ends are achieved by a unique DNA/DNA recognition mode orchestrated by the transposase. A 4-nt sequence, the so-called 'guide' sequence just 5′ of the foot of each IP hairpin, is bound near the active site, and through base pairing interactions, recognizes bases that are just 5′ of the cleavage site at both left and right IS ends (35). This mode of site-specific DNA recognition also allows the transposase to carry out strict site-specific cleavage and integration, the hallmarks of the family, without encoding any sequence-specific DNA binding domains.

The RAYT clade shares all the key amino acid motifs exhibited by IS*200*/IS*605* family TnpAs: the HUH motif and a catalytic tyrosine, as well as high sequence similarity (Figure 1b). *Escherichia coli* REPs also resemble IS*200*/IS*605* family IP sequences in that they are approximately the same length and contain mismatched bases within the hairpin stems (and hence are 'imperfect' stem-loops). REPs carry a conserved tetranucleotide (GTAG) 5′ to the hairpin foot, similar to the guide sequences of the IS*200*/IS*605* family ISs, and, like the guide sequence are proposed to be involved in cleavage specificity (26). *In vitro* assays show the conserved tetranucleotide in the presence of REP IP sequences to be required for cleavage of ssDNA containing a CT dinucleotide site. In addition, $TnpA_{REP}$ is capable of catalyzing a strand transfer reaction, one of the steps that would be required to recombine BIMEs (27).

Despite the prevalence of REPs in bacterial chromosomes, the mechanism through which they are propagated remains unclear. To better understand the functional relationship between REPs and their associated RAYTs, we determined the 3D structure of a representative from *E. coli,* $TnpA_{REP}$, to 2.6 Å in complex with a REP sequence and its conserved 5′ tetranucleotide.

MATERIALS AND METHODS

Protein expression and purification

Cloning of the $tnpA_{rep}$ gene from *E. coli* MG1655 was described previously (27). For expression of $TnpA_{REP}$ without a C-terminal His-tag, pBAD-$tnpA_{rep}$ was modified by addition of a stop codon following serine 165 to recreate wild-type sequence via site-directed mutagenesis (called pBAD-$tnpA_{rep}$-notag). Top10 cells (Novagen) were transformed with pBAD-$tnpA_{rep}$-notag. An initial overnight inoculant was grown in LB broth supplemented with 0.5% glucose, and then added to 2 l of LB broth at a 1:20 dilution and grown to an A600 nm ~0.4 at 42°C. The temperature was then dropped to 18°C, and $TnpA_{REP}$ expression induced at A600 nm 0.6 with 0.04% arabinose. Cells were harvested by centrifugation after 18 h, and resuspended in heparin binding buffer (20 mM NaH_2PO_4 pH 7.0, 500 mM NaCl, 1 mM TCEP). All subsequent steps were performed at 4°C. Lysis was by sonication. The soluble fraction was isolated by centrifugation at 13 000 rpm on a Beckman Coulter Avanti J-20 XP, loaded onto a HiTrap Heparin HP column (GE Healthcare) equilibrated in heparin binding buffer, and eluted using a linear gradient with elution buffer (20 mM NaH_2PO_4 pH 7.4, 2 M NaCl, 1 mM TCEP). The eluted protein was loaded onto a HiTrap Chelating column (GE Healthcare) pre-equilibrated with $NiSO_4$, and eluted using a linear gradient with elution buffer 2 (20 mM NaH_2PO_4 pH 7.4, 500 mM NaCl, 500 mM Imidazole, 1 mM TCEP). $TnpA_{REP}$, with DNA substrate added at a 1:1 molar ratio, was dialyzed overnight in DNA binding buffer (50 mM Tris pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 1 mM EDTA, 0.5 mM TCEP). The resulting protein–DNA complex was loaded on a HiLoad 16/60 Superdex 200 sizing column (GE Healthcare) equilibrated with DNA binding buffer. The eluted protein–DNA complex was concentrated to 5 mg/ml for crystallization trials.

DNA substrate preparation

All DNA oligonucleotides were from Integrated DNA Technologies, Inc. The DNA oligonucleotides were resuspended in 10 mM Tris pH 8 and annealed by heating to 95°C for 15 min, then rapidly cooled on ice.

Binding assay

To test binding of $TnpA_{REP}$ protein and various DNA substrates, DNA was added to the protein following elution from the HiTrap Chelating column at a 1:1 molar ratio. This was then dialyzed overnight at 4°C in

DNA binding buffer, or in DNA binding buffer with additional NaCl (Supplementary Figure S1). These mixtures were then loaded on a Superdex 200 3.2/30 column (GE Healthcare) and eluted using the same DNA binding buffer. The fractionated samples were analyzed by SDS–PAGE, and visualized by silver staining, followed by coomassie staining.

Cleavage assay

For the cleavage assay $TnpA_{REP}$ (53 μM) and DNA substrate were added together at a 1:1 molar ratio and dialyzed overnight at 4°C in DNA cleavage buffer (50 mM Tris pH 7.5, 50 mM NaCl, 1 mM EDTA, 0.5 mM TCEP). Samples were incubated at 37°C for 45 min with various divalent metal ions at 5 mM concentration. The reactions were stopped by addition of EDTA (final concentration 5 mM). The products were analyzed by SDS–PAGE.

Sedimentation velocity

$TnpA_{REP}$ was dialyzed against 50 mM Tris pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 0.5 mM EDTA and 1 mM TCEP and analyzed by sedimentation velocity at 6.7 and 13.7 μM. Sedimentation velocity experiments were conducted at 20.0°C on a Beckman Coulter ProteomeLab XL-I analytical ultracentrifuge. A total of 400 μl of each sample were loaded in two-channel centerpiece cells and analyzed at a rotor speed of 50 krpm with data collected using both the absorbance and Rayleigh interference optical detection systems. For the latter, data were collected as single scans at 280 nm using a radial spacing of 0.003 cm. Both absorbance and interference data were individually analyzed in SEDFIT12.1 b (39) in terms of a continuous c(*s*) distribution of Lamm equation solutions using an uncorrected *s* range of 0.0–5.0 S with a resolution of 100 and a confidence level of 0.68. In all cases, excellent fits were obtained with absorbance and interference RMSD values >0.0043 A_{280} and 0.0062 fringes, respectively. Solution densities *r* and viscosities *h*, and protein partial specific volumes *v* were calculated in SEDNTERP 1.09 (40).

Crystallization

Crystals of $TnpA_{REP}$ bound to the first REP in the 5′ BIME (Figure 1a) (GTAGGACGGATAAGGCGTTTA CGCCGCATCCG) were grown in hanging drops at 20°C containing a mixture of 1 μl of protein–DNA complex at 5 mg/ml with an equal volume of a reservoir solution of 4–10% PEG 5000 MME, 0.1 M MES pH 6.5 and 4–12% 1-propanol. $TnpA_{REP}$-REP crystals had $P2_12_12$ symmetry and contained one monomer in the asymmetric unit. Crystals were derivatized by soaking in 4–10% PEG 5000 MME, 0.1 M MES pH 6.5, 4–12% 1-propanol and 1 mM ethylmercury thiosalicyclic acid for 16 h.

Data collection, structure determination and refinement

$TnpA_{REP}$-REP crystals were cryoprotected using 80% mother liquor/20% glycerol mixture and flash cooled in liquid nitrogen. All diffraction data were collected at 95 K using Cu Kα radiation from a rotating anode source with multilayer focusing optics and an RAXIS IV image plate. The data were integrated and scaled with XDS (41). The structure was determined by single-wavelength anomalous diffraction from two Hg atoms. SHELXD was used to locate them (42), and their parameters were refined with SHARP (43). The map was solvent flattened with Solomon (44), and the model was built interactively with the program O (45,46). The structure was refined using CNS (47) with Cartesian simulated annealing, energy minimization and individual B factor refinement. The final model was refined with Refmac5, using restrained refinement (48). Refinement was monitored by calculating R_{free} using 5% of the data set aside for crossvalidation (49). The final model was refined to an R of 22% and an R_{free} of 28%. Drawings were prepared with PyMOL, ESPript and Adobe Illustrator (50,51). Additional X-ray data sets were collected on a single $TnpA_{REP}$-REP crystal above the K absorption edge of Fe^{2+} (7200 eV) and Mn^{2+} (6700 eV), and below Mn^{2+} (6300 eV). Using another crystal two more data sets were collected at 8200 eV and at 8500 eV, which are below and above the K edge of Ni^{2+}. These data were collected at the SERCAT beamline ID22 of the APS at the Argonne National Laboratory on a MAR300 CCD detector.

Particle induced X-ray emission

The identity of the metal ion co-factor present in $TnpA_{REP}$ was confirmed by Particle Induced X-ray Emission (PIXE) analysis at the Hope College Ion Beam Analysis Laboratory. Targets were prepared as described previously (52), where 3 μl of $TnpA_{REP}$ protein solution as well as 3 μl of blank buffer solution were dropcast and dried on separate thin aluminized mylar foil targets. Each of these targets was exposed to four 10-min irradiations with a 3.4-MeV beam of protons for a total of 0.093 nC total charge on target. The resultant X-rays were detected at 145° to the beam in a calibrated Si(Li) detector with a thin foil filter designed to shield low-energy X-rays. The measured X-ray spectra are shown in Supplementary Figure S3. The spectra are plotted on a semi-log scale and are very similar except for two elements: sulfur that arises from the cysteine and methionine groups in the protein and an unambiguous nickel K_a peak ~7.5 keV. Despite an extensive contribution of Cl from the Tris–HCl and –NaCl buffer solutions the only metal visible in the protein is nickel.

Radiolabeled reactions in vitro

Cloning and expression of the *tnpA*$_{rep}$ gene from *E. coli* MG1655 was described previously (27). $TnpA_{REP}\Delta152$ was prepared by performing site-directed mutagenesis on pBAD-*tnpA*$_{rep}$ by inverse PCR using primers His CATCA TCATCATCATCATTAAGAAG and Trp3. This was expressed and purified by growing Rosetta cells transformed with mutant plasmid at 37°C in LB broth containing carbenicillin and 1% glucose overnight. The cells were then centrifuged and diluted 50-fold into the 250 ml preheated LB medium at 37°C. Protein expression was induced at A600 nm ~0.5–0.6 by addition of arabinose to 0.04% final. After 3 h, the bacteria were centrifuged and the

pellet was washed in 10 ml of cold TN solution (50 mM Tris 7.5 100 mM NaCl) and stored at −20°C until use. The pellet then was resuspended in buffer A [50 mM Na phosphate buffer 0.1 M Na_2HPO_4 pH 8, 1 M NaCl, 10 mM β-mercaptoethanol] + 10 mM Imidazole supplemented with 1 mg/ml lysozyme and EDTA-free protease inhibitor cocktail (Roche). After 30-min incubation on ice, the bacteria were sonicated, the lysate was cleared by centrifugation and the supernatant was then mixed with Ni-agarose resin (Qiagen). After washes in buffer A + 50 mM Imidazole, $TnpA_{REP}\Delta152$ was eluted with buffer A + 200 mM Imidazole. The protein was then dialyzed in 25 mM HEPES pH 7.5, 400 mM NaCl, 1 mM EDTA, 5 mM DTT and 20% glycerol and stored at −80°C.

Oligonucleotides, purchased from Sigma and Eurogentec, were 5′-end-labeled with [γ-^{32}P] ATP (Perkin Elmer) using T4 polynucleotide kinase (NEB) or 3′-end-labeled with [α-^{32}P] dATP Cordycepin (Perkin Elmer) using Terminal Transferase (NEB), and subsequently were purified on a G25 column (GE Healthcare). Double stranded DNA was prepared by hybridization of complementary oligonucleotides. After 10 min denaturation at 98°C, the mixture was left to slowly cool to 25°C. For cleavage, 0.02 μM 5′-labeled oligonucleotide and 0.5 μM unlabeled oligonucleotide were incubated with 4 μM $TnpA_{REP}$ (45 min, 37°C, final volume 10 μl) in 12.5 mM Tris pH 7.5, 120 mM NaCl, 5 mM $MnCl_2$, 1 mM DTT, 20 μg/ml BSA, 0.5 μg of poly-dIdC and 7% glycerol. Reactions were separated on a 9% denaturing gel and analyzed by phosphor imaging.

In reactions to detect covalent complex formation, 3′ labeled substrates were incubated with $TnpA_{REP}$ in the reaction mixture as described above and reaction products were separated on a 16% SDS–PAGE gel and analyzed by phosphor imaging.

RESULTS

Overall structure

The structure of $TnpA_{REP}$ from *E. coli* MG1655 in complex with a 32-mer oligonucleotide (Figure 2a) representing a REP and its associated 5′ GTAG tetranucleotide (from the left-most BIME in Figure 1a; shown schematically in Figure 2b), was solved to 2.6 Å using single-wavelength anomalous diffraction from a mercury derivatized crystal, and refined to an R/R_{free} of 22%/28% (Table 1). The $TnpA_{REP}$ structure shows a monomer bound to one DNA molecule, consistent with the behavior of the protein in analytical ultracentrifugation analysis (Figure 2c). The DNA is bound to the protein through two extensive sets of interactions, one involving the bases at the bottom 3′ region of the IP stem, and the other the 5′ GTAG sequence. We also observe a divalent metal ion (most likely Ni^{2+}; see below) bound at the active site.

The overall fold of $TnpA_{REP}$ conforms to that of the RNA recognition motif of other HUH endonucleases (31,53,54), where a four-stranded antiparallel β-sheet is sandwiched by two helices on one side (αB and αC), and a singular helix (αD) on the other side (Figure 2a). Its topology is also very similar to those of the IS*200*/IS*605* family transposases (29,55) (Figure 1b), which is in agreement with their phylogenetic relationship (26,27).

$TnpA_{REP}$ is a monomer

Despite the clear structural and sequence similarities between $TnpA_{REP}$ and IS*200*/IS*605* family transposases at the monomer level, IS*200*/IS*605* transposases are always observed as interwoven obligatory dimers, both in the unbound form and when complexed with a variety of DNA molecules (Figure 3a and b). The complexes also always contained two IPs, binding one IP per monomer (29,55).

The comparison between $TnpA_{REP}$ and the IS*200*/IS*605* transposases reveals that two key structural determinants responsible for the dimerization interface of ∼2500 Å^2 in IS*200*/IS*605* family members are missing. As shown in Figure 3a the two β-sheets from individual IS*200*/IS*605* monomers are linked by backbone hydrogen bonding between each β5 (residues 111–115) and β2 (residues 12–18). This interaction begins with the side chain of the highly conserved T115 in β5 bonded to K18 in β2. $TnpA_{REP}$ β1, equivalent to β2 in $TnpA_{IS608}$ (Figure 1b), is markedly shorter and hence unable to be shared between monomers (Figure 3a). Furthermore, the highly conserved T117 at the start of this linkage does not exist in $TnpA_{REP}$, but is replaced by A102.

The second important structural element contributing to dimerization of IS*200*/IS*605* family transposases is a complementary hydrophobic surface at the dimer interface formed by residues contributed by both monomers. A key residue in this pocket is P73 located just after β4 in $TnpA_{IS608}$ (Figure 3b) (PDB: 2A6M, 2VHG, 2VJU) (29,35). This residue is conserved among IS*200*/IS*605* family transposases, but replaced with the polar residue E68 in $TnpA_{REP}$. Other residues that form the complementary hydrophobic surface in $TnpA_{IS608}$, such as V77 and I113, are replaced with polar residues in $TnpA_{REP}$ (S74 and E100) (Figure 1b). Although in $TnpA_{IS608}$ there are some hydrophobic interactions between the domain-swapped αD and the protein (Figure 3a), the position of this helix varies in the different structures suggesting that these interactions are not critical for dimerization.

DNA binding interactions

REP DNA is bound to $TnpA_{REP}$ through a substantial interface, which involves both the DNA hairpin and the 5′ GATG tetranucleotide; the total binding interaction between the protein and DNA buries a surface area of ∼1300 Å^2. The hairpin consists of 10 Watson–Crick base pairs, interrupted in the middle by four bases, which form a mismatched bulge (C_{27}:A_{12}, A_{13}:G_{26}), and two unpaired Ts (T_{19} and T_{20}) in the hairpin loop (bases are numbered as shown in Figure 2b). Most of the interaction involves a network of hydrogen bonds between residues comprising a region of strong positive electrostatic potential on the protein surface and the negatively charged phosphate oxygens that form the backbone of the hairpin (Figure 4a). The 5 bp closest to the hairpin tip do not contact $TnpA_{REP}$ nor do the two T bases of the tip itself.

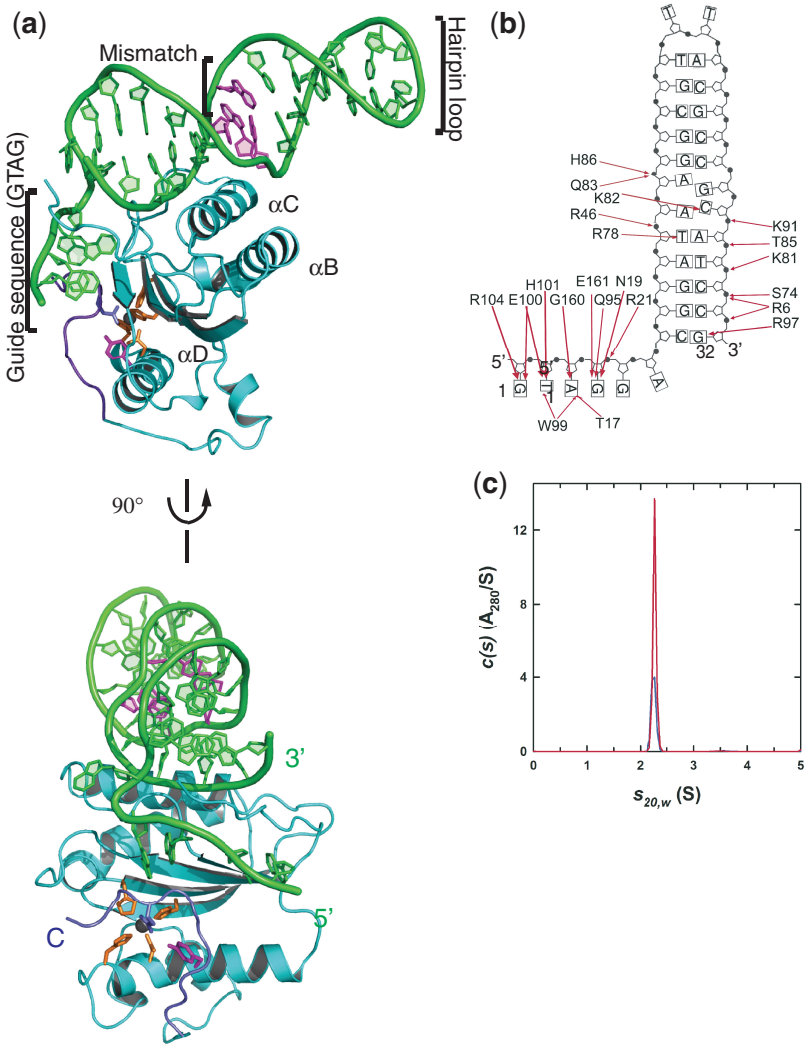

Figure 2. $TnpA_{REP}$ structure. (**a**) Ribbon diagrams of $TnpA_{REP}$ bound to hairpin, with one orientation rotated 90° around the *y*-axis relative to the other. DNA is colored green, with the bases of the mismatched bulge in magenta. The C-terminus is highlighted in purple, the catalytic tyrosine as a stick model in magenta, resides involved in the divalent metal coordination as stick models in orange and the divalent metal as a black sphere. (**b**) Schematic of the binding hairpin with important binding residues, and interactions between the DNA and protein by red arrows. (**c**) Graphical representation of the analytical ultracentrifugation by sedimentation velocity results. Sedimentation velocity experiments carried out at two loading concentrations indicated the presence of a major species at 2.35 ± 0.01 S, representing >98% of the loading signal. The best-fit molar mass for this species is 20.3 ± 0.8 kDa demonstrating that $TnpA_{REP}$ is a monomer at these concentrations.

The observed interactions suggest that binding of the hairpin is mediated mostly, but not exclusively, through the secondary structure of the DNA hairpin. This observation is reinforced by hairpin binding experiments, which show a salt concentration dependency of DNA binding (Supplementary Figure S1). The positive surface charge of the protein is comprised primarily of residues from αB and αC (R46, R78, K81, K82, Q83, T85, H86, K91 and R97) with αC inserted into the minor groove of the hairpin. Interestingly, there are base specific interactions between K82 and the mismatched C_{27}, R78 and T_{11}, and R97 and G_{32} (Figure 2b). In $TnpA_{REP}$, the K82 interaction is critical for overall binding, since mutation of A_{12} to G and A_{13} to C to correct the mismatch abolishes hairpin binding as judged by a loss of co-migration of protein and DNA by size exclusion chromatography (Figure 4b, substrate designated 'REP plus GTAG no bulge') (27). In addition, randomizing the nucleotides in the hairpin, while maintaining the hairpin structure and sequence of the bulge, also abrogates binding (Figure 4b: 'REP plus GTAG random'). These results confirm the importance of these three protein–base specific interactions.

In sharp contrast to the interactions formed with the hairpin, binding of the conserved 5′ GTAG sequence is highly base-specific and buries ~590 $Å^2$ (Figure 5a). The four bases are splayed across the surface of the protein, pointing inward such that only the phosphate backbone is accessible to solvent. This differs from the structure of $TnpA_{IS608}$ bound to its IP hairpin, which was extended

Table 1. Data collection and refinement statistics

	$TnpA_{REP}$ at 8027 eV	$TnpA_{REP}$ at 8200 eV	$TnpA_{REP}$ at 8500 eV
Data collection			
Space group	$P2_12_12$	$P2_1$	$P2_1$
Cell dimensions			
a, *b*, *c* (Å)	100.83, 60.30, 70.78	39.44, 61.50, 70.20	39.52, 61.66, 70.34
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 95.12, 90.00	90.00, 95.09, 90.00
Resolution (Å)	30–2.6 (2.67–2.6)*	60–2.5 (2.57–2.6)*	60–2.5 (2.57–2.5)*
R_{sym} or R_{merge}	4.0 % (56.5%)	3.1 % (51.5%)	3.7 % (57.2%)
$I/\sigma I$	23.46 (2.23)	21.28 (2.23)	18.66 (2.06)
Completeness (%)	99.7% (100.0%)	99.6 (99.9%)	99.6% (99.8%)
Redundancy	3.5 (3.5)	3.7 (3.7)	3.7 (3.7)
Refinement			
Resolution (Å)	2.6		
No. reflections	12 056		
R_{work}/R_{free}	22%/28%		
No. atoms			
Protein	2080		
Ligand/ion	1		
Water	22		
B-factors			
Protein/DNA	54.11		
Ligand/ion	38.93		
Water	43.71		
RMS deviations			
Bond lengths (Å)	0.016		
Bond angles (°)	1.930		

*Values in parentheses are for highest-resolution shell.

to include its 5′ tetranucleotide guide sequence. In that structure, the four 5′ nucleotides at the base of the hairpin curl away from the surface of the protein (Figures 5b and 6b), apparently poised to recognize and bind the cleavage site. In particular, two bases of the GTAG sequence in $TnpA_{REP}$, G_4 and A_3, point toward the active site and form a number of H-bonds with residues of the C-terminus around E161 (Figure 5a).

In vitro binding experiments showed that the four 5′ tetranucleotide bases are crucial for REP binding. In their absence, binding to the isolated hairpin was not detected (Figure 4b: 'REP') (27). On the other hand, while this sequence is necessary for hairpin binding, it is not sufficient, as addition of the conserved sequence to an iREP does not confer detectable binding (Figure 4b: 'iREP plus GTAG').

It was also shown that the four bases comprising the 5′ GTAG sequence are crucial for cleavage activity, as changing the sequence from GTAG to ACGA results in the loss of cleavage (27). To further understand the role of the 5′ GTAG, we individually changed each of the four bases and then tested for cleavage and binding of these modified substrates. Each base change was chosen to preserve as much as possible the observed DNA protein interactions while disrupting optimal base pairing. As shown in Figure 5c, when cleavage was monitored by following the formation of a covalent intermediate, mutation of any of the four bases resulted in the loss of cleavage activity (compare lane 4 with lanes 5–8). Also, DNA binding could no longer be detected (Supplementary Figure S2). The reciprocal substitution, changing G_4 to C and the C of the cleavage site to G, did not rescue cleavage (Figure 5c; lane 13).

Active site

As shown in Figure 6a, the active site of $TnpA_{REP}$ brings together the catalytic tyrosine Y115 and the HUH motif (H59, M60, H61), which together with two residues from αD (N119, H123) and E161 of the C-terminus, octahedrally coordinate a co-purifying metal ion. When compared with the active sites of Y1 transposases, the active site of $TnpA_{REP}$ most closely resembles that of $TnpA_{ISDra2(Y132F)}$ bound to the right end of IS*Dra2* from *Deinococcus radiodurans* (PDB: 2XO6) (55). The inactivating Y132F mutation of the active site tyrosine allowed the crystallographic capture of the fully assembled active site, including the scissile phosphate (55). Superposition of 2XO6 with $TnpA_{REP}$ shows that the αD helices align, and are positioned directly over β2 and β3, and in both cases the catalytic tyrosines face into the active site (Figure 6b).

The position of the co-purifying metal ion at the $TnpA_{REP}$ active site is the same as the catalytically essential divalent cations seen in other HUH nucleases (29,35,55,56). To determine the metal in the active site, we collected X-ray diffraction data at several energies near the K absorption edges of several metal ions, including 8200 eV (below the Ni^{2+} K edge), and at 8500 eV (above the Ni^{2+} K edge). Comparisons of the peak heights in the anomalous difference Fourier maps using these data suggest the metal is most likely Ni^{2+}. The identity of the metal ion was further confirmed via Particle-induced X-ray emission (PIXE) that again indicated the presence of Ni^{2+} in a sample that was prepared identically to the one used for crystallization (Supplementary Figure S3). Ni^{2+} is bound to the HiTrap Chelating affinity column (GE Healthcare)

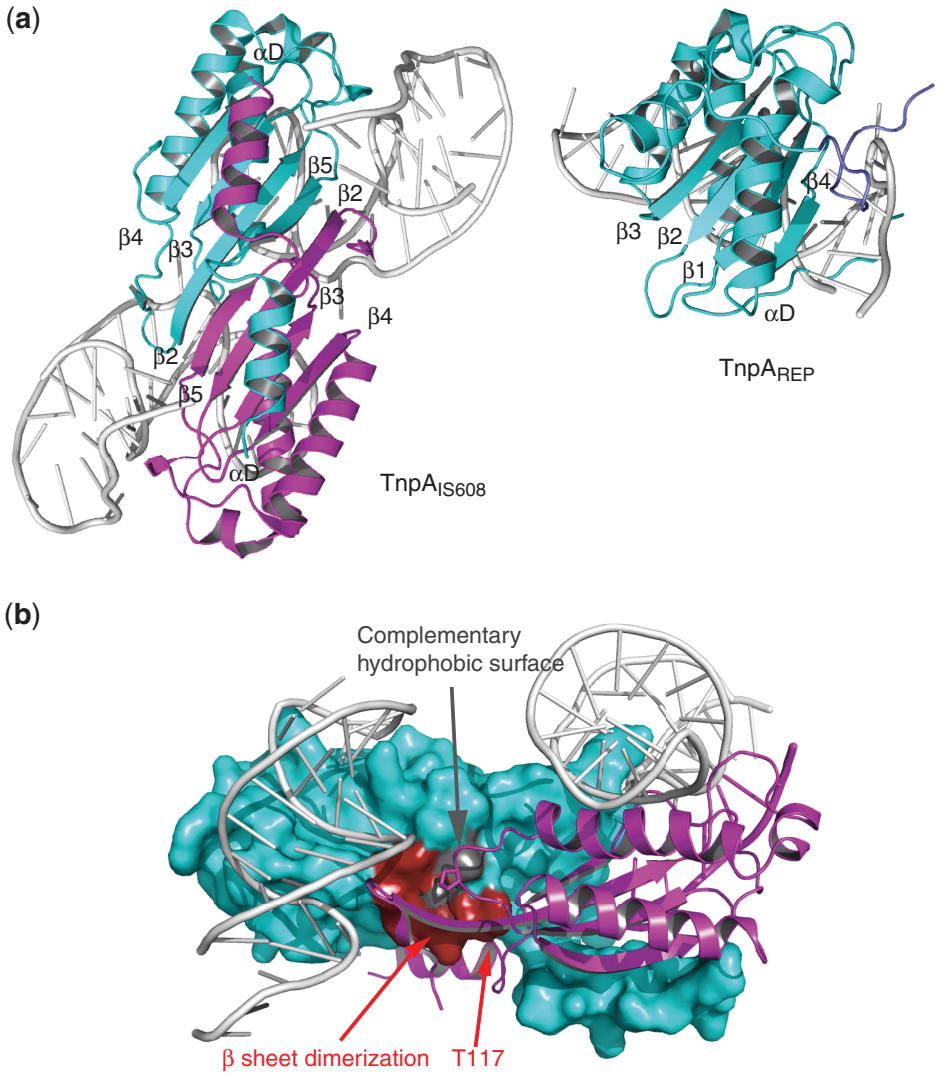

Figure 3. TnpA$_{REP}$ versus TnpA$_{IS608}$ and dimerization. (**a**) (Left) TnpA$_{IS608}$ as a ribbon diagram with one molecule colored cyan and the other molecule in magenta, with the DNA in gray. Note that αD containing the catalytic tyrosine is domain swapped. (Right) TnpA$_{REP}$ is aligned to the cyan monomer of TnpA$_{IS608}$, and is also colored in cyan, with DNA in gray. Note the differences in the positioning of αD. In each molecule the β-strands are labeled (β2, β3, β4, β5 of TnpA$_{IS608}$ are equivalent to β1, β2, β3, β4 of TnpA$_{REP}$). (**b**) One monomer of the TnpA$_{IS608}$ dimer is represented as a surface diagram in cyan and the second molecule as a ribbon diagram in magenta, with all DNA in gray. The two major components of dimerization are highlighted, with the key surface area involved in the β sheet interaction shown in red and the surface area involved in hydrophobic interaction in gray.

used during purification, and thus is the likely source of the metal observed in the structure. Significantly, we found that only Mn^{2+} supports cleavage and the formation of a covalent intermediate, while no activity was observed in the presence of other divalent cations (Figure 6c).

The most surprising aspect of the active site organization is that the carboxylate group of the metal-coordinating residue E161 occupies the same position as the scissile phosphate in the TnpA$_{Dra2(Y132F)}$ transposase crystal structure (55,57). E161 is conserved among many of the RAYT proteins, and is part of an ~25 amino acid C-terminal extension not found in the IS*200*/IS*605* transposases (Figure 1b and Figure 7b). As shown in Figure 7a, this C-terminal extension (shown in purple) packs intimately into a long surface groove that wends its way in a semicircular path across the surface of TnpA$_{REP}$.

The location of the C-terminal extension suggests that TnpA$_{REP}$ is in an auto-inhibited state in the crystal, as it appears to be physically blocking access of a ssDNA cleavage substrate to the active site. We reasoned if the C-terminus of TnpA$_{REP}$ were indeed acting as an inhibitor, one way to relieve this inhibition would be to delete it. We therefore expressed and purified a deletion mutant of TnpA$_{REP}$ where all residues after G152 had been deleted

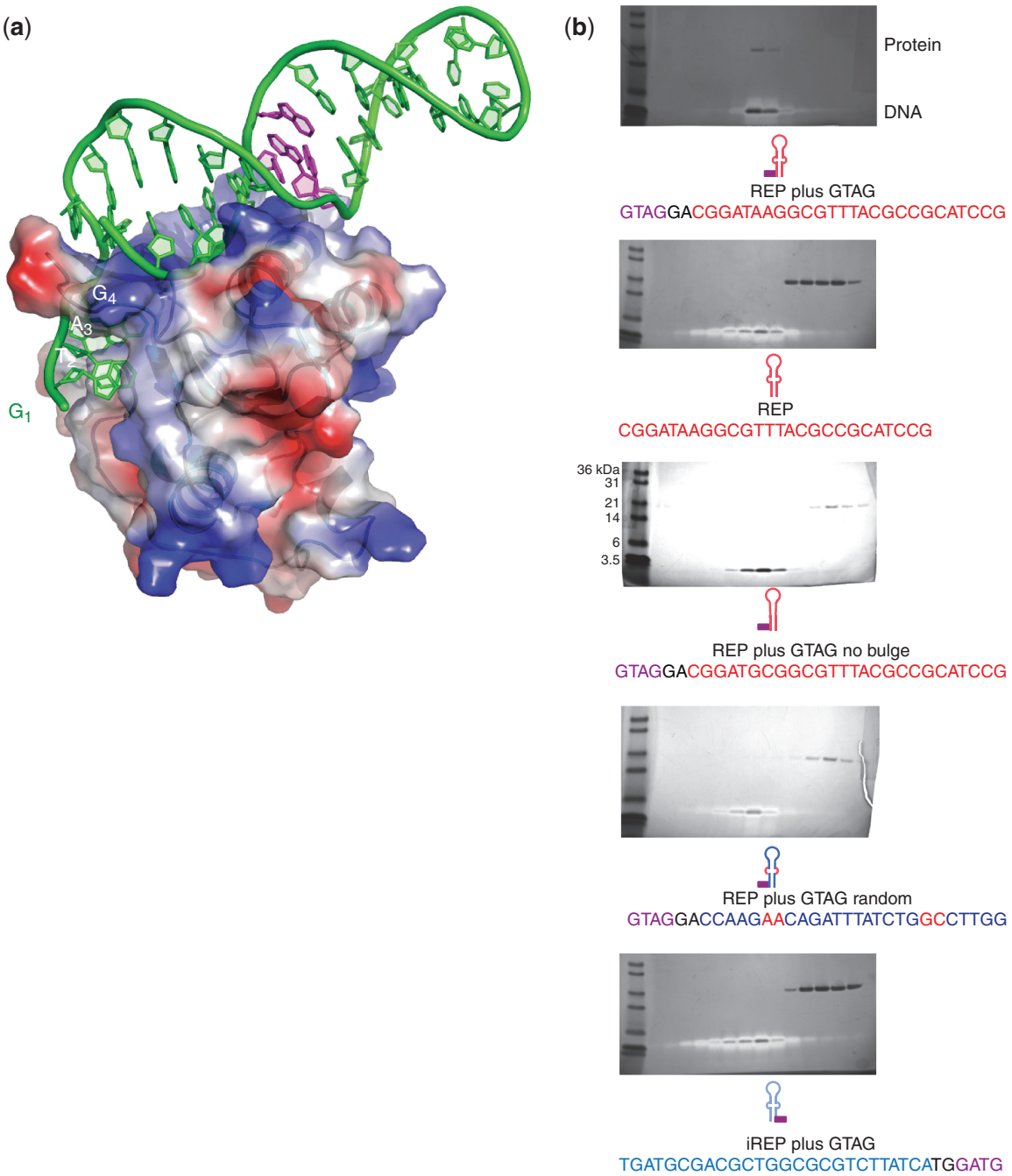

Figure 4. $TnpA_{REP}$ DNA hairpin binding. (**a**) View of the $TnpA_{REP}$ molecule showing electrostatic charge surface in a vacuum bound to its hairpin. Shades of blue represent positive charge, and red negative charge. The DNA is colored green with the individual bases of the 5′ GTAG labeled. (**b**) SDS–PAGE analysis of $TnpA_{REP}$ binding by size exclusion chromatography of modified REP and iREP substrates. The lanes in each gel represent the same elution volume off of a Superdex 200 3.2/30 column (GE Healthcare). The protein and DNA were visualized by successive staining with silver and coomassie blue. The top gel represents the wild-type binding between protein and DNA substrate (i.e. co-migration of protein and DNA), while the four subsequent gels show a lack of DNA binding as evidenced by lack of co-migration.

(designated $TnpA_{REP}\Delta152$), and tested it for cleavage activity. As shown in Figure 7c, C-terminal truncation of $TnpA_{REP}$ indeed resulted in a increase of cleavage activity relative to the full-length protein. $TnpA_{REP}\Delta152$ was also competent for strand recombination (Supplementary Figure S4), indicating that the C-terminus is not needed for strand transfer.

Cleavage site recognition

It was previously reported that $TnpA_{REP}$ is specific for cleavage at CT dinucleotide sequences, and that the CT can be located on either side of the REP hairpin and at a considerable distance from the REP hairpin (27). In addition, the inhibitory C-terminal extension of

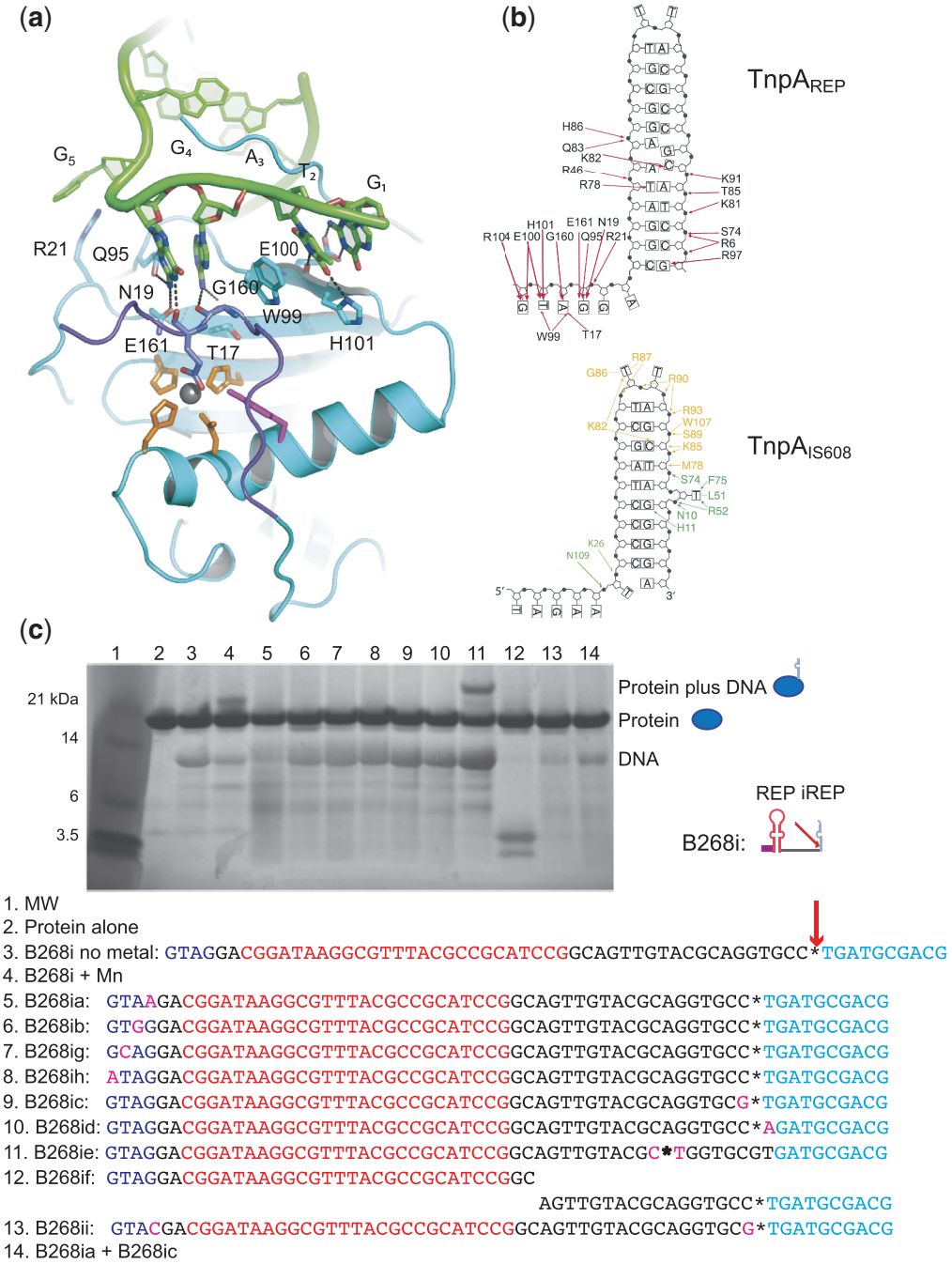

Figure 5. Analysis of the 5′ GTAG sequence. (**a**) Ribbon diagram of $TnpA_{REP}$ in cyan with key residues and bases involved in guide sequence binding shown as stick models. Hydrogen bonds are shown as dashed lines. (**b**) Schematic display of the binding hairpin from $TnpA_{REP}$ with important binding residues as in Figure 2b, with the corresponding schematic display of the binding hairpin with binding residues from $TnpA_{IS608}$. (**c**) In the top part is displayed a typical SDS–PAGE analysis of $TnpA_{REP}$ cleavage in which the DNA and protein are denatured. The DNA is visualized by silver stain, and the protein is visualized by coomassie stain. To the right of the gel are labels defining the bands, with Protein plus DNA label emphasizing protein–DNA covalent complex in the higher bands of lanes 4 and 11. A schematic of substrate used is to the right. Below is a table of substrates tested for cleavage by $TnpA_{REP}$, where the red arrow/asterisk indicates the cleavage site, the purple lettering the guide sequence, the red lettering the hairpin, the blue lettering iREP sequence and magenta lettering nucleotides that have been mutated.

$TnpA_{REP}$ makes a number of hydrogen bonds with two bases of the 5′ GTAG sequence, G_4 and A_3, which point toward the active site (Figure 5a). In the IS*200*/IS*605* transposase family, the guide sequence bases recognize nucleotides just 5′ of the cleavage site and therefore determine cleavage specificity. Assuming a similar role for G_4 and A_3 in $TnpA_{REP}$ would also imply cleavage site specificity at a CT.

The observed conformation of the 5′ GTAG sequence is remarkable. A_3 and G_4 are stacked on each other and in

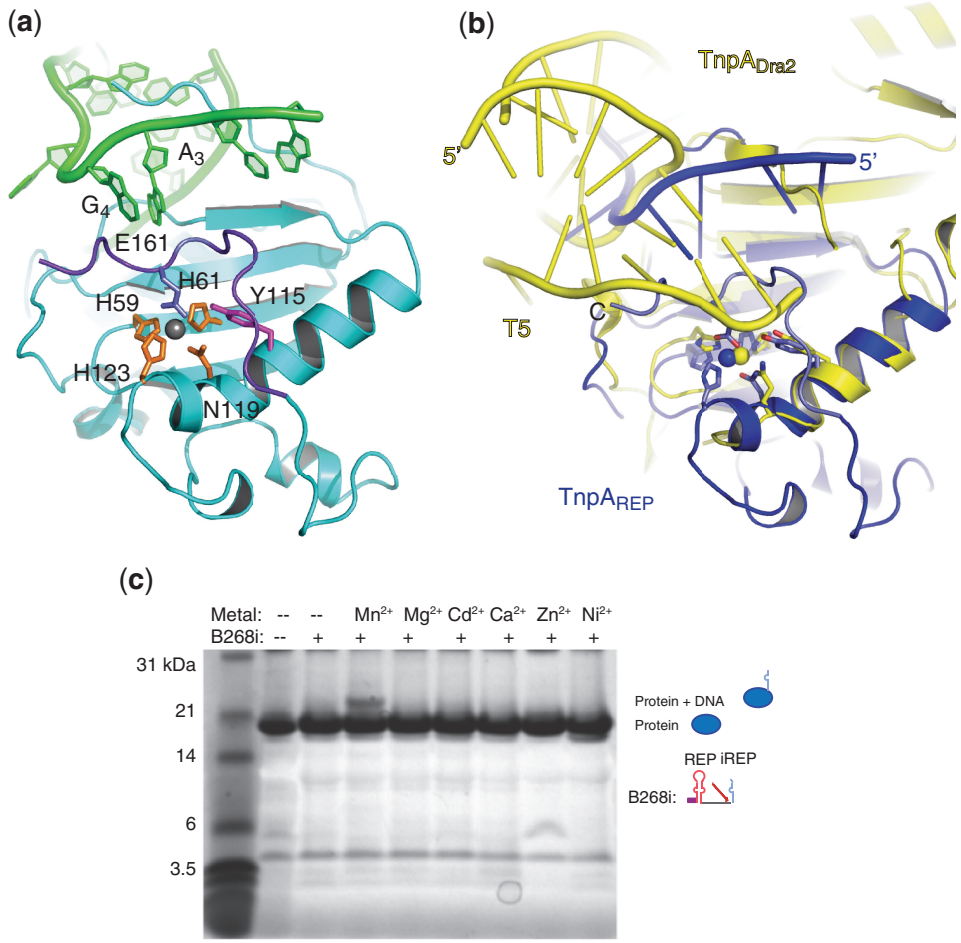

Figure 6. $TnpA_{REP}$ active site. (**a**) The active site of $TnpA_{REP}$ with critical residues highlighted. Protein/DNA is displayed as a ribbon diagram, with the protein in cyan and DNA in green. Key ligands to the cation, colored dark gray, are displayed in orange (H59, H61, N119, H123) and purple (E161). The C-terminus is highlighted in purple. (**b**) $TnpA_{REP}$ and DNA in blue and $TnpA_{Dra2}$ and DNA in yellow shown as ribbon diagrams to highlight the T5 substrate of $TnpA_{Dra2}$ and the C-terminus of $TnpA_{REP}$, and differences in 5′ GTAG conformation. Key residues in the active site of $TnpA_{REP}$ and $TnpA_{Dra2}$ are colored blue and yellow, respectively, with their oxygens colored red and nitrogens blue. (**c**) A typical SDS-PAGE analysis of $TnpA_{REP}$ cleavage in which the DNA and protein are denatured. The DNA is visualized by silver stain, and the protein is visualized by coomassie stain. To the right of the gel are labels defining the bands, with Protein plus DNA label emphasizing protein-DNA covalent complex in the higher bands of lanes 4. A schematic of substrate used is to the right.

principle can base pair with the CT of a cleavage substrate. However, in the crystal structure, the Watson–Crick face of A_3 is pointing away from the active site and the nucleotide is in the anti conformation. G_4 is also in the anti conformation but its Watson–Crick face points toward the active site and therefore is available for base-pairing. For A_3 to become available for base-pairing, it would either have to flip into the syn conformation, or, more likely, some DNA backbone movement would be necessary to turn the base so it can recognize the T of the cleavage site sequence. Notably, none of the nucleotides flanking A_3 and G_4 are available for base-pairing neither with cleavage site bases nor with any of the adjacent bases. W99 is positioned between A_3 and T_2, moving $T_2 \sim 9$ Å away from the active site. Neither T_2 nor G_1 is available for base-pairing, because of extensive hydrogen bonding with the protein. Similarly, due to the insertion of R21 and Q95, G_5 is ~12 Å away, and while its Watson–Crick face is partially accessible, it is also held in syn due to a hydrogen bond to a non-bridging oxygen of its own phosphate. Both its distance from G_4 and the geometry of the backbone makes it very unlikely that G_5 could participate in cleavage site recognition. Taken together, it appears that the observed mode of 5′ GTAG binding assures that only A_3 and G_4 are available for substrate recognition, consistent with the CT dinucleotide sequence requirement for cleavage.

As shown in Figure 5c (lanes 9, 10), mutation of the C or T at the cleavage site to G and A, respectively, abolishes activity. In an artificial substrate in which we moved the CT dinucleotide upstream, closer to the hairpin by eight bases, cleavage was moved to the new location of the CT (Figure 5c; lane 11). This result confirmed that the CT is necessary and sufficient to determine the cleavage site. Interestingly, if the enzyme and substrate were working *in trans*, as was the case for both $TnpA_{IS608}$ and $TnpA_{Dra2}$, then the combination of substrates 5 and 9 would provide one correct cleavage site in substrate 5 and

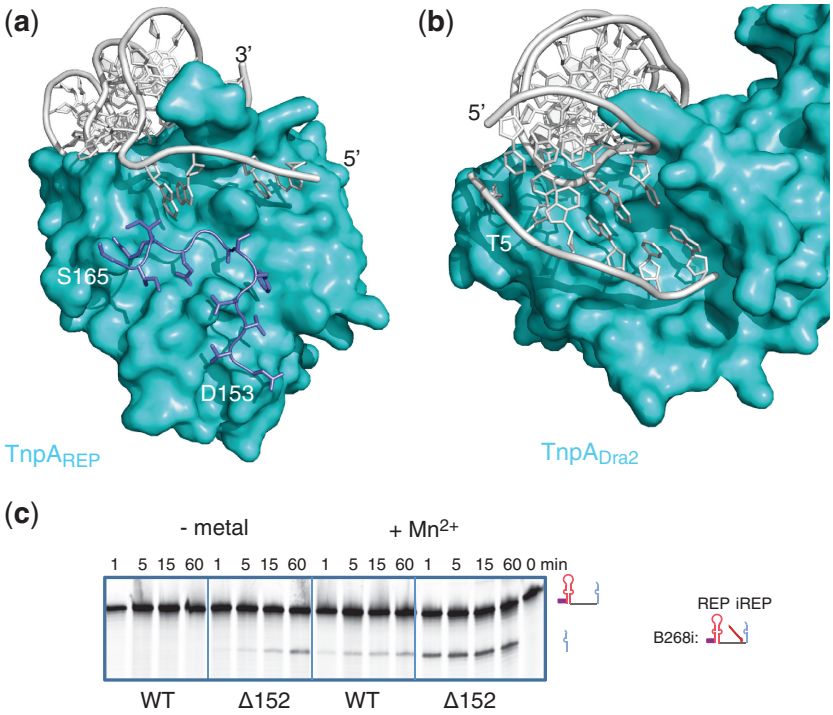

Figure 7. TnpA$_{REP}$Δ152 and the role of the C-terminus. (**a**) TnpA$_{REP}$ protein shown with molecule surface modeled in cyan, DNA in light gray as a ribbon diagram and the C-terminus as a stick model colored in purple. (**b**) TnpA$_{Dra2}$ also as a surface model in cyan, with DNA in light gray in the same orientation as TnpA$_{REP}$ in panel a to highlight the similar role and position of the TnpA$_{REP}$ C-terminus to the T5 DNA substrate of TnpA$_{Dra2}$. (**c**) SDS–PAGE analysis of DNA cleavage by TnpA$_{REP}$ and TnpA$_{REP}$Δ152.

one correct 5′ GTAG in substrate 9 from each protein/DNA complex (Figure 5c; lane 14). However, this did not rescue activity, suggesting no *in trans* cooperation between complexes.

DISCUSSION

Repetitive extragenic repeat sequences (REPs) are found in many bacterial species in high copy numbers and they can have a variety of functions (21–24). These widespread REPs are associated with TnpA$_{REP}$, closely related to Y1 transposases of the IS*200*/IS*605* family (26,27), and there is strong evidence suggesting that they may be capable of acting upon or perhaps even mobilizing REPs. Excision of REPs was reported in *P. fluorescens* (28), and it was recently demonstrated that the TnpA$_{REP}$ from *E. coli* strain MG1655 is capable of ssDNA cleavage and recombination (27). The structure of TnpA$_{REP}$ suggests that although many of the hallmarks of Y1 transposases are retained, the enzyme is highly regulated and exhibits structural features important for keeping its activity in check.

The structure of TnpA$_{REP}$ bound to a REP hairpin including its conserved 5′ tetranucleotide extension we observed is evocative of the structures of the IS*200*/IS*605* transposases bound to their DNA intermediates. Together with the previously determined structures of IS*608* and IS*Dra2* transposases it provides an elegant explanation for the CT specificity of cleavage by TnpA$_{REP}$. In particular, the pattern of base recognition first observed for IS*608* (Figure 8a) (35) and later echoed in IS*Dra2* (Figure 8b) (55) follows the same rules of base–base interaction. Thus G$_4$ of TnpA$_{REP}$ dictates C on the 5′ side of the cleavage site, while A$_3$ dictates T on the 3′ side (Figure 8c). While this explains the role of A$_3$ and G$_4$ within the proposed guide sequence, the role of other two conserved nucleotides, G$_1$ and T$_2$, is less clear. This raises the question of whether the conserved 5′ GTAG guide sequence is dictating cleavage specificity in the same fashion as the IS*200*/IS*605* transposases. As shown in the schematic in Figure 8, in IS*200*/IS*605* transposases, three of the four guide sequence nucleotides interact with nucleotides 5′ of the cleavage sequence to dictate cleavage after specific tetra- or pentanucleotide sequences (35). In contrast, only two bases constitute the cleavage site of TnpA$_{REP}$ with a C on the 5′ side and T on the 3′ site of the strand break.

Our results are consistent with this different mode of base recognition, and the crystal structure indicates that the 5′ GTAG sequence is bound in such a way that only 2 nt, A$_3$ and G$_4$, are available for base-pairing with the cleavage substrate sequence. Specific protein residues (e.g. W99, R21, Q95) assure that flanking nucleotides are sequestered and unavailable, thereby preventing them from contributing as specificity determinants. This is consistent with analysis of cleavage in *E. coli* MG1655 and *P. fluorescens* (27,28) (Figure 5c) where cleavage occurs at CT sites irrespective of the flanking nucleotides.

It is intriguing that there is considerable flexibility in the location of the CT dinucleotide sequence relative to the REP hairpin (27). The reliance only on 2 nt for specificity

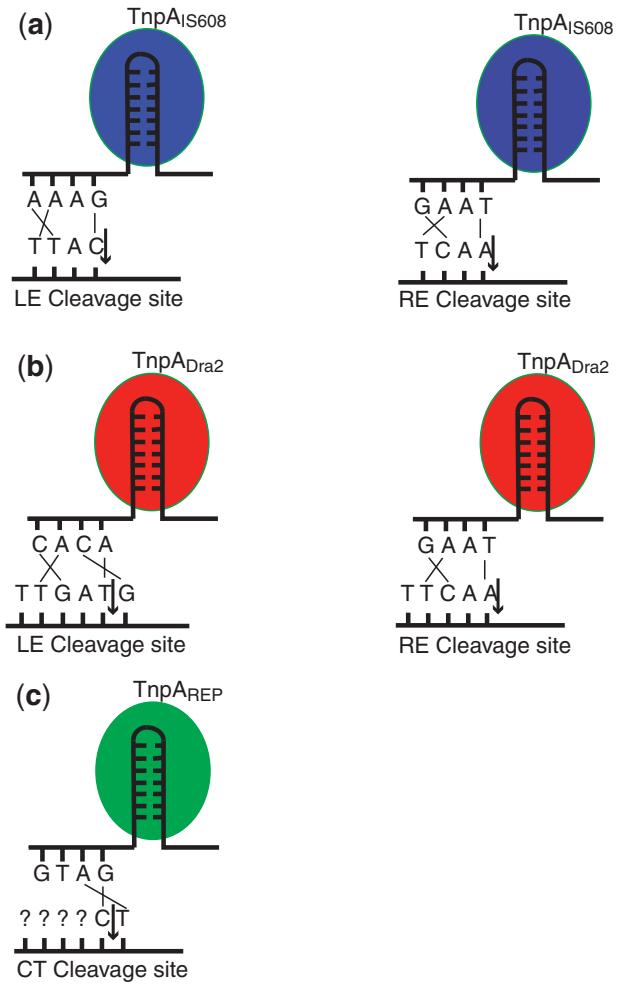

Figure 8. Models of guide sequence cleavage site recognition. (**a**) Scheme of $TnpA_{IS608}$ guide sequence cleavage site recognition. (**b**) Scheme of $TnpA_{Dra2}$ guide sequence cleavage site recognition. (**c**) Scheme of $TnpA_{REP}$ guide sequence cleavage site recognition.

and binding may imply that the exact geometry of scissile phosphate presentation for cleavage is less precise than seen in the IS*200*/IS*605* transposases. Correspondingly, we observe an active site in $TnpA_{REP}$ that prefers Mn^{2+} rather than Mg^{2+}, most likely due to three coordinating imidazoles (H59, H61, H123). A search of the MIPS database (58) reveals that there are only eight structures currently in the PDB where Mg^{2+} is coordinated by three imidazoles (and, in some cases, close inspection of these coordinate sets raises suspicions about the assigned metal's identity) but 109 structures where Mn^{2+} is coordinated by three histidine ligands. As the geometry of the octahedral coordination around a Mn^{2+} ion can be less strict than that of Mg^{2+} (59), Mn^{2+} might be more suitable for $TnpA_{REP}$ as it has to deal with a range of cleavage substrates which may not be precisely positioned within the active site. Interestingly, conjugative relaxases, which also bind their ssDNA cleavage substrates without stabilizing base pairing interactions, similarly use three histidines to coordinate a metal ion cofactor, although a preference for Mn^{2+} has not been demonstrated for this class of enzyme (32,60–62).

Molecular machines carrying out transposition reactions are often found in forms that have suboptimal activity. This is understandable as highly active runaway transposition may cause genomic damage and could result in the loss of viability. In turn, this gives rise to the possibility of engineering hyperactive versions of both prokaryotic and eukaryotic transposases for use in a variety of genomic applications (63–65). It is now also clear that stress conditions such ionizing radiation can sometimes adventitiously relieve transposon inhibition (66–68), perhaps as a last resort in an attempt to generate genetic diversity and speed up adaptation. In the structure of $TnpA_{REP}$ bound to REP DNA, the C-terminal tail is bound within the active site suggesting that this conserved sequence feature of $TnpA_{REP}$ proteins is a mechanism to inhibit or down-regulate the potential activity of the protein. Consistent with this notion, C-terminal truncation results in dramatically increased cleavage activity, while retaining competence for strand recombination (Figure 7c). Thus, this C-terminal region is not important for catalysis but instead appears to serve a regulatory role.

The $TnpA_{REP}$ structure shows a novel way in which down-regulation of a DNA rearranging system can be achieved by using a part of the transposase to act as an inhibitor of its own active site. While autoinhibition is a regulatory feature of many enzyme systems, and autoregulation has been observed through *in vitro* study of the DNA transposases Tn5, Tn10 and IS911 (69–72), to our knowledge this is the first time that it has been seen in a structure of a DNA transposase working as competitive autoinhibitor. It would be interesting to establish if there is a signal or particular growth condition in *E. coli* that activates $TnpA_{REP}$ either by interaction with other cellular proteins, by proteolytic removal of the inhibiting C-terminal tail or perhaps by the production of a truncated and hence hyperactive form of $TnpA_{REP}$ such as the deletion version we characterized.

Another consequence of removal of the C-terminal tail from the active site is that this would necessarily destroy the interactions observed between G_4 and A_3 of the conserved 5′ tetranucleotide and the residues around E161 (Figure 5a). This could provide yet another level of regulation as, at least prior to binding a cleavage substrate, the REP 5′ GTAG sequence is bound to $TnpA_{REP}$ through a dense network of interactions that renders it apparently unavailable to recognize a suitable cleavage site. The C-terminal tail binding appears to be very tight as, despite extensive efforts, we have not been able to detect binding of DNA containing a CT cleavage site (added in the form of an oligonucleotide) to a pre-formed $TnpA_{REP}$–REP complex. This has, to date, prevented us from structurally characterizing a cleavage site complex using the type of experimental approaches that proved successful for the IS*608* and IS*Dra2* transposases.

One of the most intriguing aspects of the known biochemical activities of $TnpA_{REP}$ is its apparent ability to carry out strand transfer *in vitro*. Transposases of the

IS*200*/IS*605* family can do this straightforwardly as they are obligatory dimers with two active sites and can bind simultaneously both left and right ends of the mobile element (73). Following end cleavages, the 3′ end of the cleaved strands stays in the active site, bound there by base-pairing interactions with the protein-bound guide sequences. The other end becomes covalently attached to the transposase through a 5′-phosphotyrosine linkage to the nucleophilic tyrosine located on a mobile helix. A conformational change, in which the two helices carrying the nucleophilic tyrosine swap places within the dimer, delivers the attached 5′ ends to the other active site of the dimer where the parked 3′-OH of the cleaved strand can then attack the 5′-phosphotyrosine. This reestablishes the phosphodiester backbone with a different connectivity and hence strands are transferred. It is not obvious how this could happen in the context of the monomeric $TnpA_{REP}$.

The organization of the $TnpA_{REP}$ active site when bound to the GTAG sequence suggests one possible mechanism for strand transfer. It seems likely that after cleavage at a CT dinucleotide, the ssDNA 5′ of the cleavage site would readily dissociate. As the covalent 5′-phosphotyrosine linkage remains, in principle any other ssDNA possessing a 3′-OH end could enter the active site and chemistry could proceed, thereby resolving the 5′-phosphotyrosine linkage and accomplishing strand transfer. The notion of a monomeric HUH nuclease catalyzing a strand transfer reaction, while not common, is not novel. The conjugative relaxase TrwC of plasmid R388 (but interestingly not the related TraI of the F plasmid) can catalyze strand transfer (74) as can a deletion mutant of TrwC that is monomeric. Furthermore, strand transfer occurs even with the Y26F mutant of TrwC, in which only one of the two catalytic tyrosines (Y18) is functional. Interestingly, similar to $TnpA_{REP}$ the 3′ end nick site is not bound by base pairing interactions giving rise to the possibility that it can be replaced by an 3′-OH end resulting in strand transfer (32).

The key observation that $TnpA_{REP}$ is a monomer—in contrast to the characterized IS*200*/IS*605* transposases, which are obligate dimers—imposes several constraints on any model of its activities. For example, there is no ready mechanism to explain the possibility of coordinated cleavage events on the two ends of a single BIME unless these are mediated through the DNA rather than by the protein. Furthermore, it is not clear how a circular BIME intermediate could be excised from a DNA strand in the absence of the type of reciprocal strand exchange steps between two active sites that have been invoked for the IS*200*/IS*605* family transposases (35). Any proposed mechanism for propagation must also take into account that, whereas $TnpA_{REP}$ binds REPs, binding to iREPs has not been detected (27) (Figure 4b). Although we cannot rule out that $TnpA_{REP}$ undergoes a major conformational rearrangement during the reaction that could result in dimerization, we have no evidence that it does so. Indeed, our structure demonstrates that even upon REP binding, $TnpA_{REP}$ remains monomeric.

While sequence similarities and many of our structural observations point to the close relationship between $TnpA_{REP}$ and transposases from the IS*200*/IS*605* family, the overall evidence here puts into question whether or not $TnpA_{REP}$ uses an IS*200*/IS*605* transposition mechanism. In particular, the loss of the ability to dimerize, as well as the differing role for the conserved 5′ tetranucleotide sequence in the REPs, point to crucial differences in protein activities. Furthermore, the enzyme appears highly regulated through auto-inhibition by the C-terminus and use of manganese, suggesting an evolved mechanism to limit REP populations in the cell. In all probability, $TnpA_{REP}$ started as an IS*200*/IS*605* transposase, but subsequently evolved into a REP propagation enzyme developing its own distinct 'transposition' mechanism that is kept under tight check.

ACCESSION NUMBERS

Coordinates have been deposited in the Protein Data Bank with accession code 4ER8.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

We would especially like to thank John Chrzas at the SERCAT 22-ID beamline for collecting the data at 8200 eV and 8500 eV, and Megan M. Sibley, and Paul A. DeYoung, for help with the PIXE measurements, that identified the bound metal ion species. We thank B. Marty for expert technical assistance. Data were also collected at the SER-CAT 22-ID beamline at the Advance Photon Source, Argonne National Laboratory.

FUNDING

Intramural Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (F.D., in part); Nancy Nossal Fellowship award from NIDDK (to S.M.); Postdoctoral Intramural Research Training Award from NIDDK (to S.M.); CNRS (France, the work in Toulouse was supported by intramural funding); ANR [285051 to M.C.] and US Department of Energy, Basic Energy Sciences, Office of Science [Contract No. W-31-109-Eng-38, use of APS]. Funding for open access charge: Intramural Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Lee,R., Feinbaum,R. and Ambros,V. (2004) A short history of a short RNA. *Cell*, **116**, S89–92.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

3. Garneau,J.E., Dupuis,M.E., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadan,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.
4. Horvath,P. and Barrangou,R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167–170.
5. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
6. Boccard,F. and Prentki,P. (1993) Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3′ end of transcription units. *EMBO J.*, **12**, 5019–5027.
7. 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.
8. Gilson,E., Perrin,D. and Hofnung,M. (1990) DNA polymerase I and a protein complex bind specifically to E. coli palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res.*, **18**, 3941–3952.
9. Gilson,E., Rousset,J.P., Clément,J.M. and Hofnung,M. (1986) A subfamily of E coli palindromic units implicated in transcription termination? *Ann. Inst. Pasteur Microbiol.*, **137B**, 259–270.
10. 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.
11. Oggioni,M.R. and Claverys,J.P. (1999) Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in Streptococcus pneumoniae. *Microbiology*, **145**, 2647–2653.
12. Black,C.G., Fyfe,J.A. and Davies,J.K. (1995) A promoter associated with the neisserial repeat can be used to transcribe the uvrB gene from Neisseria gonorrhoeae. *J. Bacteriol.*, **177**, 1952–1958.
13. De Gregorio,E., Abrescia,C., Carlomagno,M.S. and Di Nocera,P.P. (2003) Ribonuclease III-mediated processing of specific Neisseria meningitidis mRNAs. *Biochem. J.*, **374**, 799–805.
14. Rouquette-Loughlin,C.E., Balthazar,J.T., Hill,S.A. and Shafer,W.M. (2004) Modulation of the mtrCDE-encoded efflux pump gene complex of Neisseria meningitidis due to a Correia element insertion sequence. *Mol. Microbiol.*, **54**, 731–741.
15. Gilson,E., Clément,J.M., Brutlag,D. and Hofnung,M. (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in E. coli. *EMBO J.*, **3**, 1417–1421.
16. Higgins,C.F., Ames,G.F., Barnes,W.M., Clement,J.M. and Hofnung,M. (1982) A novel intercistronic regulatory element of prokaryotic operons. *Nature*, **298**, 760–762.
17. Gilson,E., Saurin,W., Perrin,D., Bachellier,S. and Hofnung,M. (1991) Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.*, **19**, 1375–1383.
18. Gilson,E., Saurin,W., Perrin,D., Bachellier,S. and Hofnung,M. (1991) The BIME family of bacterial highly repetitive sequences. *Res. Microbiol.*, **142**, 217–222.
19. Bachellier,S., Saurin,W., Perrin,D., Hofnung,M. and Gilson,E. (1994) Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). *Mol. Microbiol.*, **12**, 61–70.
20. Bachellier,S., Clement,J.M. and Hofnung,M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol.*, **150**, 627–639.
21. 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.
22. Gilson,E., Bachellier,S., Perrin,S., Perrin,D., Grimont,P.A.D., Grimont,F. and Hofnung,M. (1990) Palindromic unit highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae. *Res. Microbiol.*, **141**, 1103–1116.
23. Tobes,R. and Ramos,J.L. (2005) REP code: defining bacterial identity in extragenic space. *Environ. Microbiol.*, **7**, 225–228.
24. Aranda-Olmedo,I., Tobes,R., Manzanera,M., Ramos,J.L. and Marqués,S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida. *Nucleic Acids Res.*, **30**, 1826–1833.
25. Louws,F.J., Bell,J., Medina-Mora,C.M., Smart,C.D., Opgenorth,D., Ishimaru,C.A., Hausbeck,M.K., de Bruijn,F.J. and Fulbright,D.W. (1998) rep-PCR-mediated genomic fingerprinting: a rapid and effective method to identify Clavibacter michiganensis. *Phytopathology*, **88**, 862–868.
26. 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.
27. Ton-Hoang,B., Siguier,P., Quentin,Y., Onillon,S., Marty,B., Fichant,G. and Chandler,M. (2011) Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. *Nucleic Acids Res*, **40**, 3596–3609.
28. 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.
29. 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.
30. Garcillan-Barcia,M.P. and de la Cruz,F. (2002) Distribution of IS91 family insertion sequences in bacterial genomes: evolutionary implications. *FEMS Microbiol. Ecol.*, **42**, 303–313.
31. Hickman,A.B., Ronning,D.R., Kotin,R.M. and Dyda,F. (2002) Structural unity among viral origin binding proteins: crystal structure of the nuclease domain of adeno-associated virus Rep. *Mol. Cell*, **10**, 327–337.
32. Guasch,A., Lucas,M., Moncalián,G., Cabezas,M., Pérez-Luque,R., Gomis-Rüth,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.
33. Datta,S., Larkin,C. and Schildbach,J.F. (2003) Structural insights into single-stranded DNA binding and cleavage by F factor TraI. *Structure*, **11**, 1369–1379.
34. Koonin,E.V. and Ilyina,T.V. (1993) Computer-assisted dissection of rolling circle DNA replication. *Biosystems*, **30**, 241–268.
35. 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.
36. 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.
37. Ton-Hoang,B., Guynet,C., Ronning,D.R., Cointin-Marty,B., Dyda,F. and Chandler,M. (2005) Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. *EMBO J.*, **24**, 3325–3338.
38. Mahillon,J. and Chandler,M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.*, **62**, 725–774.
39. Schuck,P. and Rossmanith,P. (2000) Determination of the sedimentation coefficient distribution by least-squares boundary modeling. *Biopolymers*, **54**, 328–341.
40. Cole,J.L., Lary,J.W., Moody,T.P. and Laue,T.M. (2008) Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. *Methods Cell Biol.*, **84**, 143–179.
41. Kabsch,W. (2010) Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 133–144.
42. Sheldrick,G.M. (2008) A short history of SHELX. *Acta Crystallogr. A*, **64**, 112–122.
43. Vonrhein,C., Blanc,E., Roversi,P. and Bricogne,G. (2007) Automated structure solution with autoSHARP. *Methods Mol. Biol.*, **364**, 215–230.
44. Abrahams,J.P. and Leslie,A.G.W. (1996) Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr. D Biol. Crystallogr.*, **52**, 30–42.
45. Jones,T.A. and Kjeldgaard,M. (1997) Electron-density map interpretation. *Methods Enzymol.*, **277**, 173–208.
46. Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A.*, **47**, 110–119.

47. Brünger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J., Kuszewski,J., Nilges,M., Pannu,N.S. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 905–921.
48. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.*, **53**, 240–255.
49. Brünger,A.T. (1997) Free R value: cross-validation in crystallography. *Methods Enzymol.*, **277**, 366–396.
50. Gouet,P., Courcelle,E., Stuart,D.I. and Métoz,F. (1999) ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics (Oxford, England)*, **15**, 305–308.
51. DeLano,W.L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.
52. Warner,J.D., DeYoung,P.A., Ellsworth,L.A., Kiessel,L.M., Rycenga,M.J. and Peaslee,G.F. (2010) Quantitative analysis of a metalloprotein compositional stoichiometry with PIXE and PESA. *Nucl. Instrum. Meth. B*, **268**, 1671–1675.
53. Burd,C.G. and Dreyfuss,G. (1994) Conserved structures and diversity of functions of RNA-binding proteins. *Science*, **265**, 615–621.
54. Cléry,A., Blatter,M. and Allain,F.H. (2008) RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.*, **18**, 290–298.
55. 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.
56. Larkin,C., Datta,S., Nezami,A., Dohm,J.A. and Schildbach,J.F. (2003) Crystallization and preliminary X-ray characterization of the relaxase domain of F factor TraI. *Acta Crystallogr. D Biol. Crystallogr.*, **59**, 1514–1516.
57. Larkin,C., Datta,S., Harley,M.J., Anderson,B.J., Ebie,A., Hargreaves,V. and Schildbach,J.F. (2005) Inter- and intramolecular determinants of the specificity of single-stranded DNA binding and cleavage by the F factor relaxase. *Structure*, **13**, 1533–1544.
58. Hemavathi,K., Kalaivani,M., Udayakumar,A., Sowmiya,G., Jeyakanthan,J. and Sekar,K. (2010) MIPS: metal interactions in protein structures. *J. Appl. Crystallogr.*, **43**, 196–199.
59. Harding,M.M. (1999) The geometry of metal-ligand interactions relevant to proteins. *Acta Crystallogr. D Biol. Crystallogr.*, **55**, 1432–1443.
60. Boer,R., Russi,S., Guasch,A., Lucas,M., Blanco,A.G., Pérez-Luque,R., Coll,M. and de la Cruz,F. (2006) Unveiling the molecular mechanism of a conjugative relaxase: the structure of TrwC complexed with a 27-mer DNA comprising the recognition hairpin and the cleavage site. *J. Mol. Biol.*, **358**, 857–869.
61. Lucas,M., González-Pérez,B., Cabezas,M., Moncalian,G., Rivas,G. and de la Cruz,F. (2010) Relaxase DNA binding and cleavage are two distinguishable steps in conjugative DNA processing that involve different sequence elements of the nic site. *J. Biol. Chem.*, **285**, 8918–8926.
62. Monzingo,A.F., Ozburn,A., Xia,S., Meyer,R.J. and Robertus,J.D. (2007) The structure of the minimal relaxase domain of MobA at 2.1 Å resolution. *J. Mol. Biol.*, **366**, 165–178.
63. Goryshin,I.Y. and Reznikoff,W.S. (1998) Tn5 in vitro transposition. *J. Biol. Chem.*, **273**, 7367–7374.
64. Mátés,L., Chuah,M.K.L., Belay,E., Jerchow,B., Manoj,N., Acosta-Sanchez,A., Grzela,D.P., Schmitt,A., Becker,K., Matrai,J. *et al.* (2009) Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.*, **41**, 753–761.
65. Yusa,K., Zhou,L., Li,M.A., Bradley,A. and Craig,N.L. (2011) A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl Acad. Sci. USA*, **108**, 1531–1536.
66. Twiss,E., Coros,A.M., Tavakoli,N.P. and Derbyshire,K.M. (2005) Transposition is modulated by a diverse set of host factors in Escherichia coli and is stimulated by nutritional stress. *Mol. Microbiol.*, **57**, 1593–1607.
67. Pasternak,C., Ton-Hoang,B., Coste,G., Bailone,A., Chandler,M. and Sommer,S. (2010) Irradiation-induced Deinococcus radiodurans genome fragmentation triggers transposition of a single resident insertion sequence. *PLoS Genet.*, **6**, e1000799.
68. Dai,J., Xie,W., Brady,T.L., Gao,J. and Voytas,D.F. (2007) Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin. *Mol. Cell*, **27**, 289–299.
69. Mahnke Braam,L.A., Goryshin,I.Y. and Reznikoff,W.S. (1999) A mechanism for Tn5 inhibition. Carboxyl-terminal dimerization. *J. Biol. Chem.*, **274**, 86–92.
70. Allingham,J.S. and Haniford,D.B. (2002) Mechanisms of metal ion action in Tn10 transposition. *J. Mol. Biol.*, **319**, 53–65.
71. Allingham,J.S., Wardle,S.J. and Haniford,D.B. (2001) Determinants for hairpin formation in Tn10 transposition. *EMBO J.*, **20**, 2931–2942.
72. Duval-Valentin,G. and Chandler,M. (2011) Cotranslational control of DNA transposition: a window of opportunity. *Mol. Cell*, **44**, 989–996.
73. Montaño,S.P. and Rice,P.A. (2011) Moving DNA around: DNA transposition and retroviral integration. *Curr. Opin. Struct. Biol.*, **21**, 370–378.
74. César,C.E., Machón,C., de la Cruz,F. and Llosa,M. (2006) A new domain of conjugative relaxase TrwC responsible for efficient oriT-specific recombination on minimal target sequences. *Mol. Microbiol.*, **62**, 984–996.