

UC San Diego

UC San Diego Previously Published Works

Title

Diversity-generating Retroelements in Phage and Bacterial Genomes

Permalink

<https://escholarship.org/uc/item/9tf9w0sd>

Journal

Microbiology Spectrum, 2(6)

ISSN

2165-0497

Authors

Guo, Huatao
Arambula, Diego
Ghosh, Partho
et al.

Publication Date

2014-11-21

DOI

10.1128/microbiolspec.mdna3-0029-2014

Peer reviewed

Diversity-generating Retroelements in Phage and Bacterial Genomes

HUATAO GUO,¹ DIEGO ARAMBULA,²
PARTHO GHOSH,³ and JEFF F. MILLER^{2,4}

¹Department of Molecular Microbiology and Immunology, University of Missouri School of Medicine Columbia, MO 65212; ²Department of Microbiology, Immunology and Molecular Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095; ³Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093; ⁴The California NanoSystems Institute, University of California at Los Angeles, Los Angeles, CA 90095

ABSTRACT Diversity-generating retroelements (DGRs) are DNA diversification machines found in diverse bacterial and bacteriophage genomes that accelerate the evolution of ligand–receptor interactions. Diversification results from a unidirectional transfer of sequence information from an invariant template repeat (TR) to a variable repeat (VR) located in a protein-encoding gene. Information transfer is coupled to site-specific mutagenesis in a process called mutagenic homing, which occurs through an RNA intermediate and is catalyzed by a unique, DGR-encoded reverse transcriptase that converts adenine residues in the TR into random nucleotides in the VR. In the prototype DGR found in the *Bordetella* bacteriophage BPP-1, the variable protein Mtd is responsible for phage receptor recognition. VR diversification enables progeny phage to switch tropism, accelerating their adaptation to changes in sequence or availability of host cell-surface molecules for infection. Since their discovery, hundreds of DGRs have been identified, and their functions are just beginning to be understood. VR-encoded residues of many DGR-diversified proteins are displayed in the context of a C-type lectin fold, although other scaffolds, including the immunoglobulin fold, may also be used. DGR homing is postulated to occur through a specialized target DNA-primed reverse transcription mechanism that allows repeated rounds of diversification and selection, and the ability to engineer DGRs to target heterologous genes suggests applications for bioengineering. This chapter provides a comprehensive review of our current understanding of this newly discovered family of beneficial retroelements.

INTRODUCTION

Mobile genetic elements have repeatedly been called to duty in life-and-death struggles between hosts and their pathogens (1, 2, 3, 4). One of their greatest utilities is the capacity to create DNA sequence diversity in protein-encoding genes, thereby generating protective shields to defend against enemies, or to create arsenals of weapons to exploit potential hosts. After decades of research, considerable evidence now suggests that the V(D)J recombination system, which is essential for generating adaptive immunity in vertebrates, has evolved from an ancestral DNA transposon (2, 3, 4). The site-specific recombinases responsible for V(D)J recombination, RAG1 and RAG2, are able to catalyze DNA transposition in a manner analogous to DNA transposons (2), and the RAG1 core and V(D)J recombination signals are likely derived from the transposase and terminal repeats of an

Received: 30 May 2014, **Accepted:** 17 July 2014,
Published: 5 December 2014

Editors: Alan Lambowitz, University of Texas, Austin, TX, and Nancy Craig, Johns Hopkins University, Baltimore, MD

Citation: Guo H, Arambula D, Ghosh P, Miller JF. 2014. Diversity-generating retroelements in phage and bacterial genomes. *Microbiol Spectrum* 2(6):MDNA3-0029-2014. doi:10.1128/microbiolspec.MDNA3-0029-2014.

Correspondence: Jeff F. Miller, jfmiller@ucla.edu, Huatao Guo, guohua@missouri.edu

© 2014 American Society for Microbiology. All rights reserved.

ancient DNA transposon similar to *Transib* (3, 4). Ironically, pathogens also exploit mobile genetic elements to generate protein diversity, altering their antigenic characteristics to evade host immunity (1). This process of antigenic variation is employed by *Borrelia* species, *Neisseria gonorrhoeae*, and other pathogens. Bacterial antigenic variation often involves a single, highly expressed gene encoding an abundant surface protein and dozens of archived ones that are homologous but different from each other. Replacing all or part of the expressed copy by DNA transposition leads to antigenic variation on the surface of the pathogen.

Diversity-generating retroelements (DGRs) are a recently discovered class of beneficial mobile element that diversify DNA sequences and the proteins they encode (5, 6). DGRs function through a template-dependent, reverse transcriptase (RT)-mediated mechanism that introduces nucleotide substitutions at defined locations in specific genes (5, 6, 7). DGRs were initially discovered during studies of pathogenesis by *Bordetella* species, which cause respiratory diseases in humans and other animals (5). The cell surfaces of these bacteria are highly dynamic due to changes in gene expression that accompany their infectious cycles (8). In a search for transducing vectors, a group of temperate bacteriophages was discovered that possess a remarkable ability to generate tropic variants that use different cell-surface molecules for infection (5, 9). Subsequent genetic and genomic studies with the prototype phage, BPP-1, showed that tropism switching is mediated by a phage-encoded DGR. This DGR introduces nucleotide substitutions in a gene that specifies a host-cell-binding protein, which is positioned at the distal tips of phage tail fibers (Fig. 1) (5, 9). As a result, BPP-1 can adapt to dynamic changes on the surfaces of *Bordetella* species. Guided by the sequences of phage DGR components, homologous elements have been identified in numerous bacterial, plasmid, and phage genomes (6, 10, 11, 12, 13). Most DGRs are bacterial chromosomal elements and they are distributed throughout the bacterial domain, with representatives in all phyla that have significant sequence coverage. Although variations in architectures and associated components appear to mediate adaptations to particular needs, all DGRs are predicted to function in a fundamentally similar way. The BPP-1 phage serves as a model for this entire family of retroelements, and our discussion begins with a brief description of its features.

TROPISM-SWITCHING *BORDETELLA* PHAGE

Bordetella species are aerobic, Gram-negative bacterial pathogens that colonize ciliated respiratory epithelial

surfaces. *Bordetella pertussis* and *Bordetella parapertussis* are human-restricted and cause whooping cough (pertussis), while *Bordetella bronchiseptica* infects a broad range of wild and domesticated mammals (8). The infectious cycle of these closely related species is regulated by a conserved, environmentally responsive phosphorelay system composed of the BvgS sensor protein and the BvgA response regulator, which control expression of an extensive array of cell-surface and secreted molecules. In the so-called Bvg⁺ phase, BvgAS is active and induces expression of adhesins, toxins, a type III secretion system, and numerous additional factors involved in colonization of respiratory surfaces. In the Bvg⁻ phase, the BvgAS phosphorelay is suppressed, virulence genes are quiescent, and a distinct set of loci are induced. In *B. bronchiseptica*, Bvg⁻-phase genes are responsible for motility, chemotaxis, and survival under conditions of nutrient deprivation, presumably representing an *ex vivo* phase of the infectious cycle. Dynamic changes in surface molecule expression are critical for the lifestyles of *Bordetella* species, and they likely pose a challenge to infecting phage.

BPP-1 (Bvg⁺-plus tropic phage 1), isolated from a *B. bronchiseptica* strain cultured from the upper respiratory tract of a rabbit, preferentially forms plaques on Bvg⁺-phase as opposed to Bvg⁻-phase *Bordetella* (5, 9). Using a combination of deletion, complementation, and cell-binding assays, Liu *et al.* identified the receptor for BPP-1 as pertactin (Prn), an outer-membrane autotransporter protein that is included as a protective antigen in acellular pertussis vaccines (5). Expression of *prn* is activated by BvgAS, thereby explaining BPP-1 tropism. At a frequency of $\sim 10^{-5}$, however, BPP-1 formed normal-sized plaques on Bvg⁻-phase *Bordetella* (5). Since plaque formation requires repeated rounds of infection, replication, and reinfection, this simple observation suggested that a tropism switch had occurred (5). Indeed, tropic variants fell into two classes. The first, designated BMP (Bvg⁻-minus tropic phage), preferentially infected Bvg⁻-phase cells, while the second, BIP (Bvg⁻ indiscriminant phage), infected both Bvg⁺ and Bvg⁻ cells with similar efficiencies. Although these phenotypes are inheritable, tropic variants are continuously generated at characteristic frequencies by BPP, BMP, and BIP.

THE BPP-1 DGR

To identify genetic changes responsible for tropism switching, whole-genome sequences were obtained for a collection of BPP, BIP, and BMP variants (5, 9). In each case, tropism switching was accompanied by nucleotide

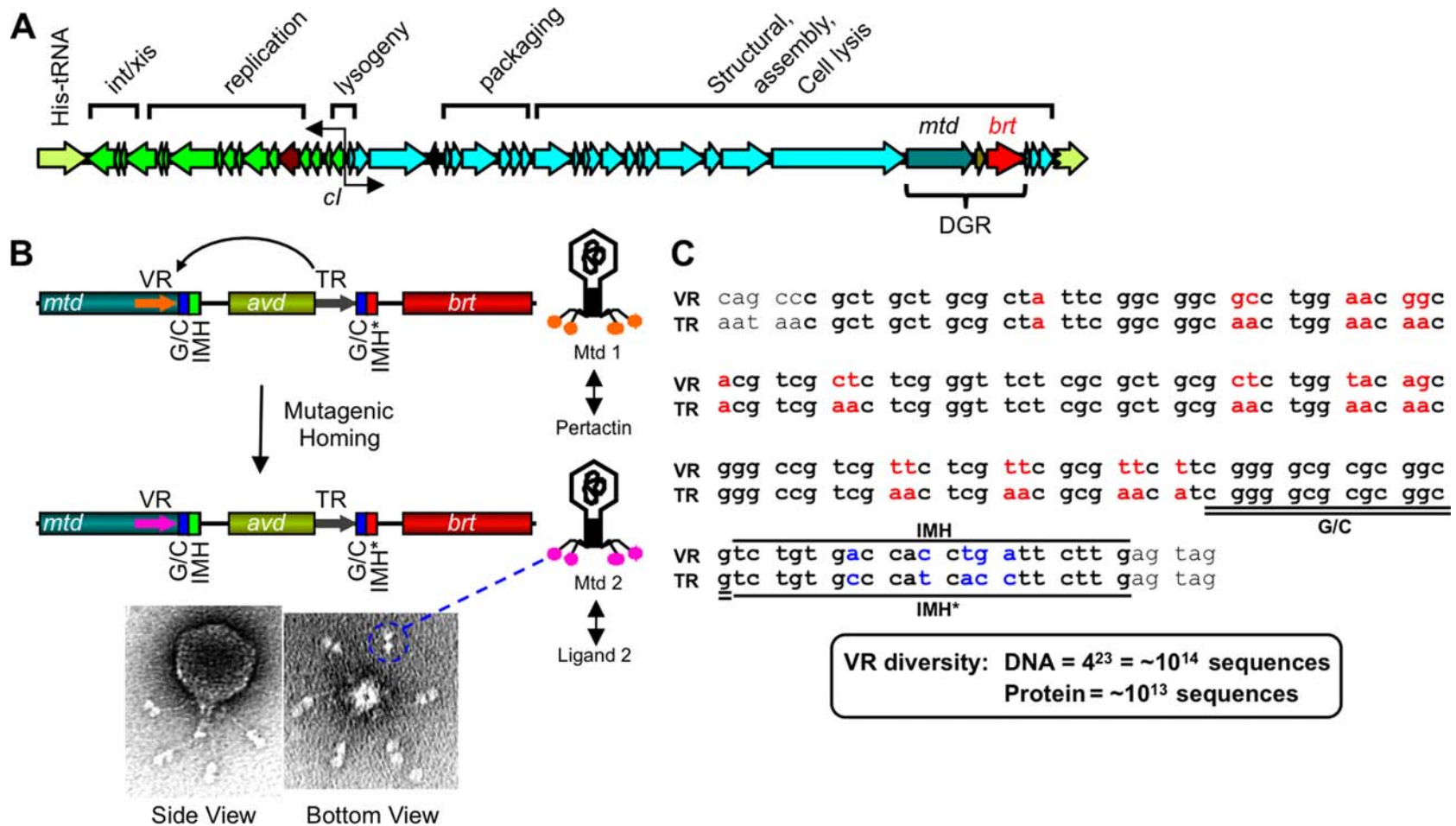


FIGURE 1 BPP-1 phage and its diversity-generating retroelement (DGR). **(A)** The BPP-1 genome is represented in the prophage form flanked by a duplication of the His-tRNA gene formed during integration. Functional assignments for most gene clusters are indicated, along with the *cl*-like repressor and the DGR cassette. **(B)** Schematic representation of the DGR cassette and its function in phage tropism switching. The cassette contains three genes (*mtd*, *avd*, and *brt*) and two 134-bp repeats (template and variable repeats, or TR and VR, respectively). The VR is located at the 3' end of the *mtd* gene, which encodes the distal tail fiber protein responsible for receptor recognition. Located at the 3' ends of the VR and TR are IMH (initiation of mutagenic homing) and IMH* elements, respectively, in addition to a GC-rich element. Phage tropism switching occurs through DGR-mediated mutagenic homing, in which TR sequence information is transferred to the VR with adenine residues in the TR appearing as random nucleotides in the VR. Shown on the bottom are electron micrographs of the BPP-1 phage; globular structures at the distal ends of tail fibers are Mtd trimers (two per fiber). **(C)** Comparison of the BPP-1 TR and VR. The TR and VR sequences are shown in bold. The VR variable positions and the corresponding adenine residues in the TR are shown in red. IMH, IMH*, and GC-rich elements are also indicated. There are 23 adenines in the TR, which can theoretically generate $\sim 10^{14}$ different DNA sequences, or $\sim 10^{13}$ different peptides. (Adapted from 7 and 9.) [doi:10.1128/microbiolspec.MDNA3-0029-2014.f1](https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014.f1)

substitutions within a 134-bp sequence, designated the variable repeat (VR), located at the 3' end of the *mtd* (major tropism determinant) gene (Fig. 1). Mtd is a trimeric tail fiber protein responsible for phage binding to *Bordetella* surfaces, and changes in its coding sequence confer new ligand specificities (14). Comparison of phage variants showed that variable nucleotides were located at a subset of positions in the VR and most often at the first two positions of codons, thereby maximizing amino acid diversity (5, 14). There are 23 variable positions at which any of the four nucleotides can be found, giving a theoretical DNA diversity of 10^{14} sequences and a resulting repertoire of ~10 trillion polypeptides. This rivals the estimated diversity that can be generated by mammalian antibody and T-cell-receptor genes (15, 16).

An imperfect repeat of the VR, designated the template repeat (TR), is found downstream of *mtd* (Fig. 1) (5). Liu *et al.* made the simple but seminal observation that variable residues in the VRs corresponded to adenine residues in the TR, which remained unchanged during phage tropism switching (5). This prompted the hypothesis that tropism switching is associated with a mutagenic mechanism that is adenine specific, and that the TR somehow provides a template for this process. Indeed, precise deletion of the BPP-1 TR resulted in fully infectious phage particles for Bvg⁺ *Bordetella* but abolished the ability to switch tropism. Furthermore, single-nucleotide substitutions introduced into the TR, which corresponded to silent mutations in the VR, appeared in the VR during tropism switching and were accompanied by mutations at adenines (5). These results demonstrated that DNA sequence information was “transferred” from the TR to the VR during tropism switching and was accompanied by adenine-specific mutagenesis.

An additional piece of the puzzle came from the unexpected finding of an RT gene, *brt* (*Bordetella* reverse transcriptase), in the double-stranded (ds)DNA BPP-1 genome (Fig. 1) (5). Its position adjacent to the VR/TR prompted the hypothesis that reverse transcription could be involved in tropism switching. To test this, an in-frame deletion was introduced into *brt* and the resulting phenotype was identical to the Δ TR deletion: infectious phage particles were produced but tropism switching was abolished. This suggested an intriguing link between reverse transcription and adenine mutagenesis. The *brt* gene encodes a 38-kDa protein with similarity to RTs found in other retroelements, including a conserved YMDD box found in RT catalytic centers. Substitution of the YMDD box with SMAA resulted in the loss of phage tropism switching *in vivo* and RT activity *in vitro*.

It was proposed that tropism switching occurs through an RT-dependent transfer of sequence information from an invariant template (the TR) to a region of variability (the VR), and is accompanied by adenine mutagenesis in a process called “mutagenic homing.” The RT dependency predicted the involvement of an RNA intermediate, leading Liu *et al.* to designate the tropism-switching cassette shown in Fig. 1 as a DGR (5). An additional gene in the BPP-1 DGR, designated *avd* (accessory variability determinant), encodes a small polypeptide (Avd) that plays an essential role in DGR function as described below (6, 9).

DGRs ARE WIDESPREAD IN BACTERIAL AND PHAGE GENOMES

The ability of the BPP-1 DGR to accelerate the evolution of novel ligand-binding specificities suggested that similar elements would be found elsewhere in nature. Not surprisingly, using *Brt* as a template, Doulatov *et al.* identified additional DGRs in phage and bacterial genomes (6). These included a *Vibrio harveyi* phage and the chromosomes of human commensals (*Bifidobacterium longum* and *Bacteroides thetaiotaomicron*), a human oral pathogen (*Treponema denticola*), and cyanobacteria. Some cyanobacterial species contained multiple DGRs, or multiple target open reading frames that are diversified *in trans* by a single DGR. DGR RTs were found to be closely related to RTs of mobile group II introns, and in all cases repeated elements corresponding to VR/TR cognate pairs could be found in nearby loci. Predicted target proteins and VR sequences were diverse, but the VRs differed from cognate TRs almost exclusively at sites corresponding to TR adenines, suggesting that DGRs function through a conserved mechanism.

More recently, Minot *et al.* performed high-throughput DNA sequencing on the gut virome of healthy humans (12). Metagenomic analysis of phage populations from stool samples identified 36 unique TR/VR pairs, 29 of which were adjacent to a DGR-type RT gene. In total, DGRs were identified in 11 out of 12 subjects studied. As with other DGRs, TR/VR sequences differed almost exclusively at sites corresponding to TR adenines. This discovery demonstrated that DGRs are common in the genomes of bacteriophages found in the lower gastrointestinal tract. Using a custom-made script called DiGReF, Schillinger *et al.* conducted a large-scale search of sequence databases and identified 155 DGRs in phage and bacteria, both Gram positive and Gram negative (13). DGRs are particularly abundant in certain phyla, with bacteroides (27.7%), firmicutes (31.0%),

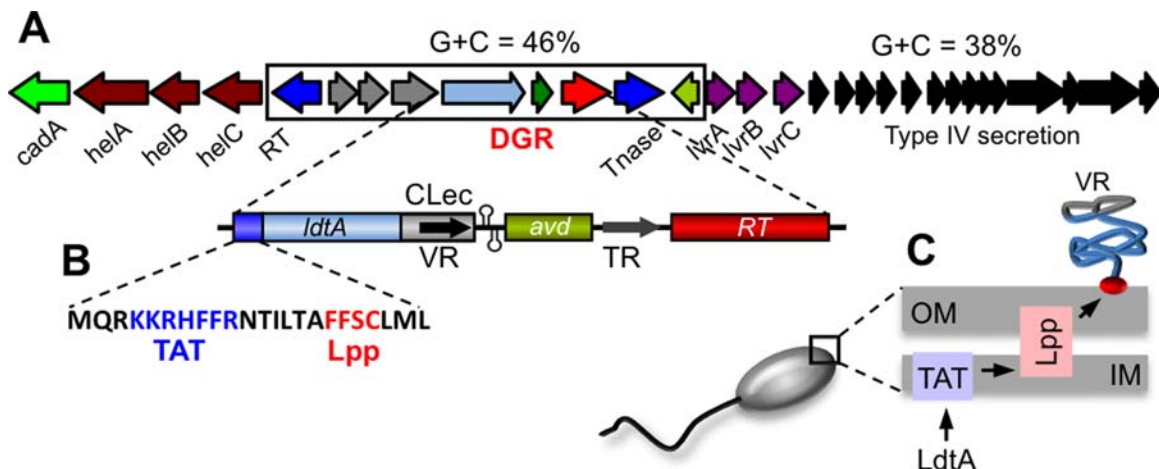
and proteobacteria (25.2%) containing the greatest numbers of unique elements. DGRs were also found in actinobacteria, cyanobacteria, *Deinococcus-Thermus*, nitrospirae, spirochaetes, chlamydiae/verrucomicrobia, and other bacterial groups. These observations indicated that DGR host organisms are highly diverse and occupy varied environmental niches. In this large dataset, nearly all the VRs were located at the 3' end of protein-coding genes and they ranged from 50 to 150 bp. The relatively short length of VRs might reflect constraints on DNA sequence hypervariation in protein-encoding genes, as highly dense nucleotide substitutions over a long stretch would be more likely to result in a loss of function. This study confirmed the earlier observation by Doulatov *et al.* that many bacterial genomes contain multiple DGRs, or encode single DGRs that diversify multiple target genes (6).

Although DGRs are often considered in the context of *Bordetella* phage, this is an artifact of their discovery and the choice of BPP-1 as a prototype for mechanistic studies. In reality, DGRs are widely distributed in bacterial chromosomes, in addition to phage and plasmid genomes.

DGR TARGET PROTEINS

The cellular localization and physiological functions of the vast majority of DGR-diversified target proteins are uncharacterized, but common themes are beginning to emerge. Arambula *et al.* recently analyzed a DGR found on the chromosome of *Legionella pneumophila*, an opportunistic pathogen (Fig. 2) (17). The *L. pneumophila* DGR is located on a conjugative transposable element and was found to be capable of mutagenic homing with characteristic adenine mutagenesis. Remarkably, its TR contains 43 adenine residues and is theoretically capable of generating 10^{26} unique DNA sequences, creating a repertoire of 10^{19} distinct proteins. The DGR-encoded target protein LdtA contains both a TAT (twin arginine transport) motif and a lipobox at its N terminus. The TAT pathway is an alternative secretion system that can translocate folded proteins or protein complexes across bacterial cytoplasmic membranes (18), and lipobox motifs mediate signal peptide cleavage, lipid modification, and anchoring to inner or outer membranes (19). Genetic and biochemical analysis showed that LdtA is indeed a TAT-secreted protein that is lipid modified and localized to the outer leaflet of the *Legionella* outer

FIGURE 2 Diversification of a surface-displayed lipoprotein by a *Legionella* DGR. **(A)** The *L. pneumophila* strain Corby DGR is encoded on a genomic island that differs in G+C content from the rest of the genome. The VR sequences at the 3' end of the diversified locus, *ldtA*, are flanked by tandem hairpin/cruciform structures that are essential for efficient homing. The TR contains 43 adenine residues, which can create a potential repertoire of 10^{26} different VR DNA sequences. **(B)** LdtA contains atypical TAT (twin arginine transport) and Lpp (lipobox, lipid modification) signals at the N terminus. **(C)** Cellular localization studies demonstrated that LdtA is exported through the inner membrane via the TAT pathway, lipid modified, and anchored on the outer surface of the outer membrane via an Lpp-like lipoprotein processing pathway. VR-encoded residues are surface displayed by a C-terminal CLec fold. (Adapted from 17.) [doi:10.1128/microbiolspec.MDNA3-0029-2014.f2](https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014.f2)



membrane, with C-terminal VR sequences exposed to the extracellular milieu (Fig. 2) (17). Bioinformatic analysis predicts that target proteins of many bacterial DGRs, including those of *T. denticola*, *Bacteroides* species, *Vibrio angustum*, and *Shewanella baltica*, also contain N-terminal lipobox motifs preceded by either TAT or Sec secretion signals (17). This suggests that lipid modification and surface display on bacterial outer membranes will be a common feature of proteins diversified by bacterial DGRs.

From both structural and functional perspectives, the BPP-1 Mtd protein is by far the most extensively characterized DGR target protein (14, 20, 21). To understand structure–function relationships, McMahon *et al.* determined the atomic structures of five different Mtd tropic variants (Fig. 3) (14). The Mtd variants formed intertwined, pyramid-shaped homotrimers with nearly identical secondary and tertiary structures, indicating that VR diversification did not lead to gross conformational changes. Each monomer was organized into three discrete domains: a β -prism, a β -sandwich, and a C-type lectin (CLec), arranged from the N to the C terminus. The β -prism domains converge to form a vertex on the top, and the CLec domains interact with each other to form the bottom part of the pyramid-shaped trimer. The VR-encoded variable amino acids are presented by CLec folds on the bottom surface of the Mtd trimer, with their side chains solvent exposed and accessible to ligand interactions. The most remarkable feature of the comparative analysis of Mtd structures is that the five distinct tropic variants showed virtually no conformational variation in the VR-encoded regions. As shown in Fig. 3, the main chain conformations of these tropic variants are nearly superimposable. This is in striking contrast to antigen-binding regions of immunoglobulin (Ig) molecules, where conformational flexibility is associated with the ability to recognize diverse antigens.

The CLec fold appears to be a conserved feature of many DGR-diversified proteins. Although DGR target proteins generally share little sequence similarity, structure-based sequence alignments predicted that VR-encoded variable residues are often presented in the context of C-terminal CLec domains (14). This was recently confirmed by X-ray crystallography using the *T. denticola* TvpA protein (Fig. 3C, D), which is predicted to be a DGR-diversified lipoprotein localized to the spirochaetal surface (22). TvpA contains a CLec fold that is highly homologous to Mtd, and VR residues are positioned in a remarkably similar manner. Instead of forming homotrimers like BPP-1 Mtd, however, TvpA

exists as a monomer and does not contain a β -prism or β -sandwich domain as found in Mtd.

Interestingly, a subset of phage DGRs were predicted to use Ig folds, similar to those found in Igs, instead of CLec folds to display variable residues (12). These Ig folds were usually located in the middle, as opposed to the 3' ends of VR-containing open reading frames. Although Igs and these predicted DGR variable proteins both use Ig folds, they appear to have evolved different means for displaying variable residues (12). In the former, diversified residues are located in flexible loop regions positioned between β -sheets. In contrast, structural modeling of Ig-type DGR target proteins indicated that diversified residues are displayed on one face of a β -sheet and the linker region connecting it to the adjacent domain. These observations suggest that DGR target proteins and antigen receptors may have evolved different solutions to accommodate sequence diversity in the context of Ig folds.

To understand better the basis of ligand recognition by DGR variable proteins, Miller *et al.* co-crystallized BPP-1 Mtd and its outer-membrane ligand, pertactin (Fig. 3E) (20). Structural analysis showed that each Mtd trimer bound to one molecule of pertactin, whose extracellular domain has an extended β -helix structure. An asymmetric mode of interaction was observed: two identical VR regions from two of the Mtd monomers in the trimer each bound a different loop from pertactin, with these loops having no sequence similarity to one another.

The fundamental basis of the evolvability of Mtd–ligand interactions was revealed, at least in part, through an analysis of binding constants (20). The K_d value for the interaction between purified Mtd trimers and the ectodomain of pertactin was 3.5 μ M as determined by surface plasmon resonance. In contrast, the K_d for the intact phage was \sim 6.9 pM, reflecting a nearly 10^6 -fold increase in binding strength. The BPP-1 phage contains six tail fibers with two Mtd trimers at their tips (21). With 12 Mtd trimers on each phage particle, and the ectodomain of pertactin displayed at high density on the *Bordetella* surface, ligand–receptor binding is multivalent and driven by avidity, which results in the exponential amplification of individual binding strengths (20). This amplification relaxes the demand for optimal complementarity between partners, greatly expanding the scope of molecules that can be recognized by DGR-diversified proteins. Avidity-driven interactions are inherently evolvable, and are predicted to characterize most, if not all, DGR variable protein–ligand interactions.

DIRECTIONALITY OF DGR MUTAGENIC HOMING

Mutagenic homing is an RT-mediated, adenine-specific, error-prone process that unidirectionally transfers sequence information from the TR to the VR (5, 6). In BPP-1, substitution of adenine residues in the TR with nonadenine nucleotides eliminated VR mutagenesis at cognate positions, while replacing nonadenine residues with adenines resulted in novel sites of mutagenesis in the VR, proving that sequence diversification is intrinsic to TR adenines.

A prominent feature of mutagenic homing is that diversity is specifically targeted to the VR while the TR sequences remain unchanged (5, 6). What determines this directionality? A comparison of the two repeats in BPP-1 revealed that, in addition to differences corresponding to adenine residues in the TR, they also differ by five base pairs at their 3' ends (6). These polymorphisms are located within a 21 bp segment downstream of a 14-bp GC-rich element common to both the TR and VR (Figs 1 and 4). During mutagenic homing, the polymorphisms are never co-converted. Swapping experiments by Doulatov *et al.* revealed that they are required for the unidirectional transfer of sequence information (6). When the 21-bp element at the 3' end of the VR was swapped with analogous sequences from the TR, the VR was no longer diversified. Conversely, replacing the 21-bp TR element with the corresponding VR sequences resulted in TR diversification at adenines, albeit at a low level. The 21-bp element in the VR was named the initiation of mutagenic homing (or IMH) element, while the corresponding sequence in TR was called IMH*. Further studies showed that similar polymorphisms could be found at the 3' ends of cognate TRs and VRs in many DGRs (6). As detailed below, DNA structural determinants that follow IMH are also required for efficient homing.

DGR HOMING OCCURS THROUGH AN RNA INTERMEDIATE

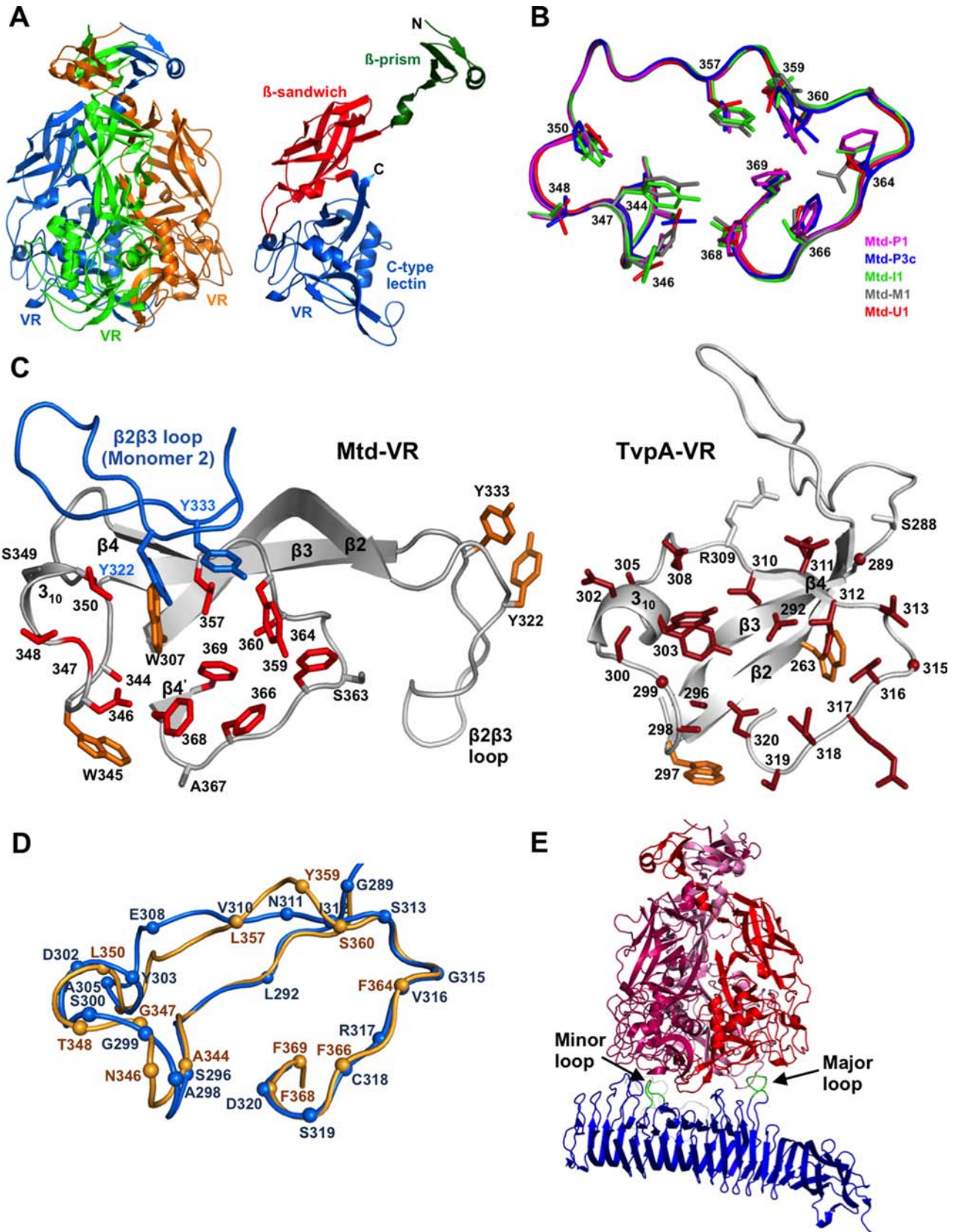
The presence of conserved RTs in DGRs and the required role of Brt in tropism switching suggested that mutagenic homing occurs through an RNA intermediate. To test this hypothesis, the BPP-1 TR was "tagged" with a self-splicing group I intron from bacteriophage T4 (the *td* intron) (7). Intron tagging was first used by Boeke *et al.* (23) to probe retrotransposition by the *Saccharomyces cerevisiae* transposon Ty1, and has subsequently been used to prove that other retroelements, including human long interspersed nuclear elements

(LINEs) and mobile group II introns, function through RNA intermediates (24, 25, 26). As intron excision takes place only at the RNA level, detection of ligated exons in DNA homing products provides genetic proof that sequence information flows from DNA to RNA to DNA.

Polymerase chain reaction (PCR)-based assays were used to detect homing products (i.e., VR sequences) derived from *td* intron-tagged TRs supplied *in trans* on a replicating plasmid (7). As predicted, homing products consisted of VR sequences that contained ligated *td* exons, and their detection required functional Brt and a splicing-competent *td* intron. Sequence analysis verified precise intron excision and adenine mutagenesis in progeny VRs. These observations conclusively showed that DGR homing had occurred through a TR RNA intermediate. Regions of the TR that are important for function were analyzed by similar PCR-based assays in which donor TRs were sequence tagged to allow detection and characterization of homing products (7). Mutational analysis demonstrated that sequences internal to the TR are largely dispensable, while sequences at either terminus are essential for the function of the RNA intermediate (7). By expressing wild-type *avd* on a compatible plasmid *in trans* to the TR and *brt*, a 300-bp portion of the *avd*-coding sequence that extends upstream of the limit of TR/VR homology was found to be required for optimal TR RNA function (27). Shorter sequences upstream of TR, of as little as 48 bp, provided partially functional TR RNA species that supported DGR homing at lower efficiencies (27). Sequences downstream of TR that are required for optimal function extend for at least 110 bp (7). Although the TR-containing RNA species that serves as an intermediate in the retrotransposition reaction has yet to be characterized, the observation that sequences that lie beyond the limits of TR/VR homology influence activity suggests that RNA stability and/or structural determinants are important for DGR-mediated homing.

cDNA SYNTHESIS AND INTEGRATION

cDNA synthesis could occur through one of several mechanisms. First, it could be initiated through a target DNA-primed RT (TPRT) reaction similar to those used by non-LTR (long terminal repeat) retroelements and mobile group II introns, which are closely related to DGRs (28, 29, 30, 31). Group II introns are site-specific retrotransposons found in bacterial, fungal, and plant organelle genomes that insert at cognate intronless alleles in a process called intron homing (32). Intron mobility is catalyzed by a ribonucleoprotein complex



consisting of both the spliced intron RNA and the intron-encoded protein with both RT and endonuclease activities. The intron RNA cleaves the DNA sense strand in an RNA-catalyzed reverse splicing reaction, while the intron-encoded protein cleaves the DNA antisense strand shortly downstream of the exon junction and uses it as a primer to reverse transcribe the RNA, leading to intron mobility. Alternative mechanisms for cDNA initiation during DGR homing include priming by an RNA moiety, a nontarget DNA, or even a protein capable of donating a free hydroxyl group. The close evolutionary relationship between RTs from DGRs and group II introns, the site-specificity of mutagenic homing, and other observations led to the hypothesis that DGRs also function through a TPRT-type mechanism (7).

To determine the site at which TR-derived cDNA integrates at the 3' end of the VR, a marker co-conversion assay was developed (7). Single-nucleotide substitutions were introduced into sequence-tagged TRs, and DGR-mediated transfer to the VR was determined by sequencing homing products amplified by PCR. A clear boundary for marker transfer was detected and localized to a 5-bp sequence within the GC element at the 3' end of the VR (Fig. 4). This could represent the site at which cDNA synthesis initiates in a TPRT reaction, or possibly the site of integration of cDNA that was primed by a non-TPRT mechanism. Marker co-conversion at the 5' end of the VR was more heterogeneous, suggesting that 5' cDNA integration can occur at different VR locations, possibly mediated by homology between VR and cDNA sequences.

The TPRT model predicts that homing occurs through sequential steps, with 3' cDNA integration taking place first (7). Indeed, when the first 99 bp of the BPP-1 VR were deleted (Δ VVR1–99) to prevent 5' cDNA integration, cDNA products linked to VR sequences at their 3' end with free, unlinked 5' ends were detected (Fig. 4B). These products were IMH and Brt dependent and contained adenine-specific mutations, suggesting that adenine mutagenesis occurs during reverse transcription and is an

intrinsic property of the DGR RT. In addition, these observations suggested that cDNA integration at the 3' end of the VR can occur independently of 5' integration, consistent with the hypothesis that DGR homing occurs through a TPRT reaction. It is important to note, however, that neither single-strand nicks nor double-strand breaks have been detected within the GC-rich element where marker co-conversion begins, and no endonuclease gene, domain, or activity has been identified in a DGR cassette, raising the possibility that DGR homing could occur through an alternative mechanism.

Is 5' cDNA integration homology dependent? To test this, Guo *et al.* inserted *mtd* sequences into the TR that were homologous to a 50-bp region immediately upstream of the deletion junction in the Δ VVR1–99 recipient (Fig. 4C) (7). This restored 5' cDNA integration and DGR homing as confirmed by PCR-based assays and sequence analysis. cDNA integration into cryptic sites was observed at low frequency, and the sites of integration mapped to short stretches of identity (4 to 12 bp) between TR–*mtd* and sequences upstream of the Δ VVR1–99 deletion. These results showed that 5' integration can be mediated by short stretches of homology between TR-derived cDNA and target DNA. One model that could account for this is template switching during reverse transcription. It has been observed previously that RTs encoded by the group II intron Ll.LtrB and other non-LTR retroelements, including the Mauriceville retroplasmid and the silkworm *Bombyx mori* R2 element, are capable of template switching during reverse transcription (33, 34, 35, 36, 37, 38, 39, 40). Interestingly, these reactions appear to occur primarily from the 5' end of the first template to the 3' end of the second template (DNA or RNA), and involve little or no base-pairing interactions between the latter and the 3' end of the cDNA. Whatever the mechanism of 5' cDNA integration might be, it appears to be RecA independent, as DGR homing and phage tropism switching occur at similar efficiencies in wild-type and RecA-deficient *Bordetella* (7).

FIGURE 3 The CLec fold as a scaffold for display of DGR-generated protein diversity. **(A)** Left: The BPP-1 Mtd protein forms a pyramid-shaped homotrimer with VR-encoded residues exposed on the bottom surface. Right: An Mtd monomer containing β -prism, β -sandwich, and VR-encoded CLec domains, from the N to the C terminus. **(B)** Backbone structures of the VR regions of five Mtd variants with different ligand specificities are shown. Despite side-chain variations in diversified VR residues, the backbone structures are nearly superimposable. **(C)** Comparison of the CLec VR regions of BPP-1 Mtd and a *T. denticola* variable protein, TvpA. For the Mtd VR, the β 2 β 3 loop of a second monomer is also shown (blue). **(D)** Superposition of the VR regions of BPP-1 Mtd (light orange) and *T. denticola* TvpA (blue). **(E)** Interaction of an Mtd homotrimer with the receptor protein pertactin. See text for details. (Adapted from 14, 20, and 22.) doi:10.1128/microbiolspec.MDNA3-0029-2014.f3

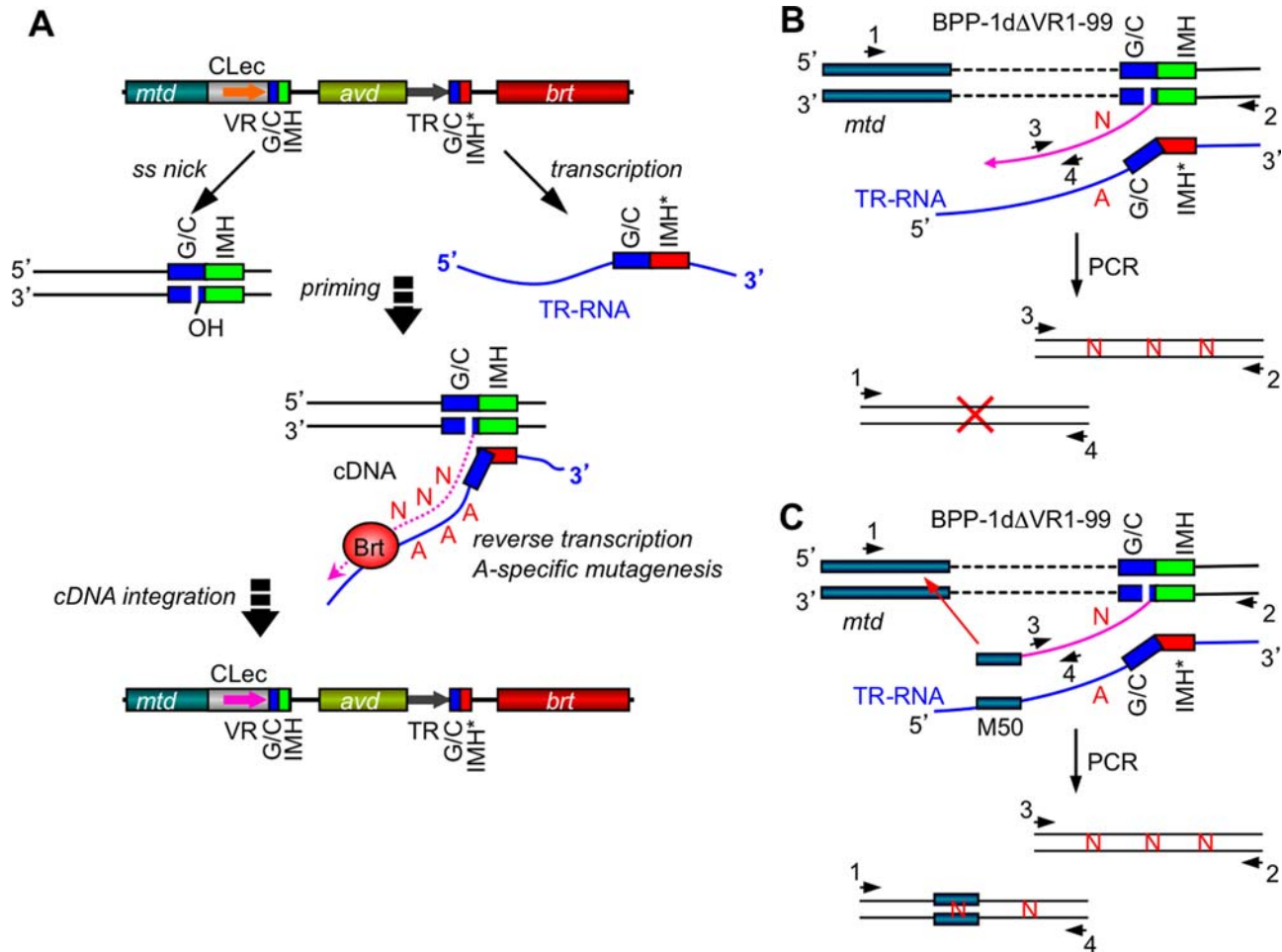


FIGURE 4 The TPRT model of BPP-1 DGR-mediated mutagenic homing. **(A)** Mutagenic homing occurs through a TR RNA intermediate and is RecA independent, similar to group II intron homing. A marker co-conversion assay mapped the cDNA transfer boundary to a narrow region within the GC-rich element at the 3' end, which may represent 3' cDNA integration site(s). The marker transfer boundary at the 5' end was more heterogeneous. A target DNA-primed reverse transcription model, similar to that of group II intron homing, has been proposed to explain these observations. The DNA target site was hypothesized to be nicked within the GC-rich element, with the exposed 3'-hydroxyl group serving as a primer for adenine-specific error-prone reverse transcription of the TR RNA. Integration of cDNA products at the 5' end requires short stretches of homology between the VR and the cDNA and may occur through strand displacement or template switching followed by break repair. Subsequent DNA replication would then create progeny genomes with mutagenized variable regions. **(B)** Deletion of the VR sequence upstream of GC and IMH elements appeared to block 5' cDNA integration but not 3' cDNA integration, as analyzed by PCR with primer sets 1/4 and 2/3, respectively. Sequence analysis showed adenine mutagenesis in PCR products generated with primers 2&3. **(C)** 5' cDNA integration in Δ VR1-99 was restored by inserting a 50 bp *mtd* sequence, which is homologous to the region upstream of the deletion junction, in the TR. (Adapted from reference 7) [doi:10.1128/microbiolspec.MDNA3-0029-2014.f4](https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014.f4)

DGR TARGET RECOGNITION

In addition to the GC-rich sequence and IMH, deletion analysis revealed a third element located downstream of the BPP-1 VR that is required for efficient homing

(Fig. 5) (41). This region contains two 8-bp GC-rich inverted repeats capable of forming a DNA hairpin or cruciform structure with a 4-nt loop. Mutations that disrupt stem formation greatly decreased DGR homing

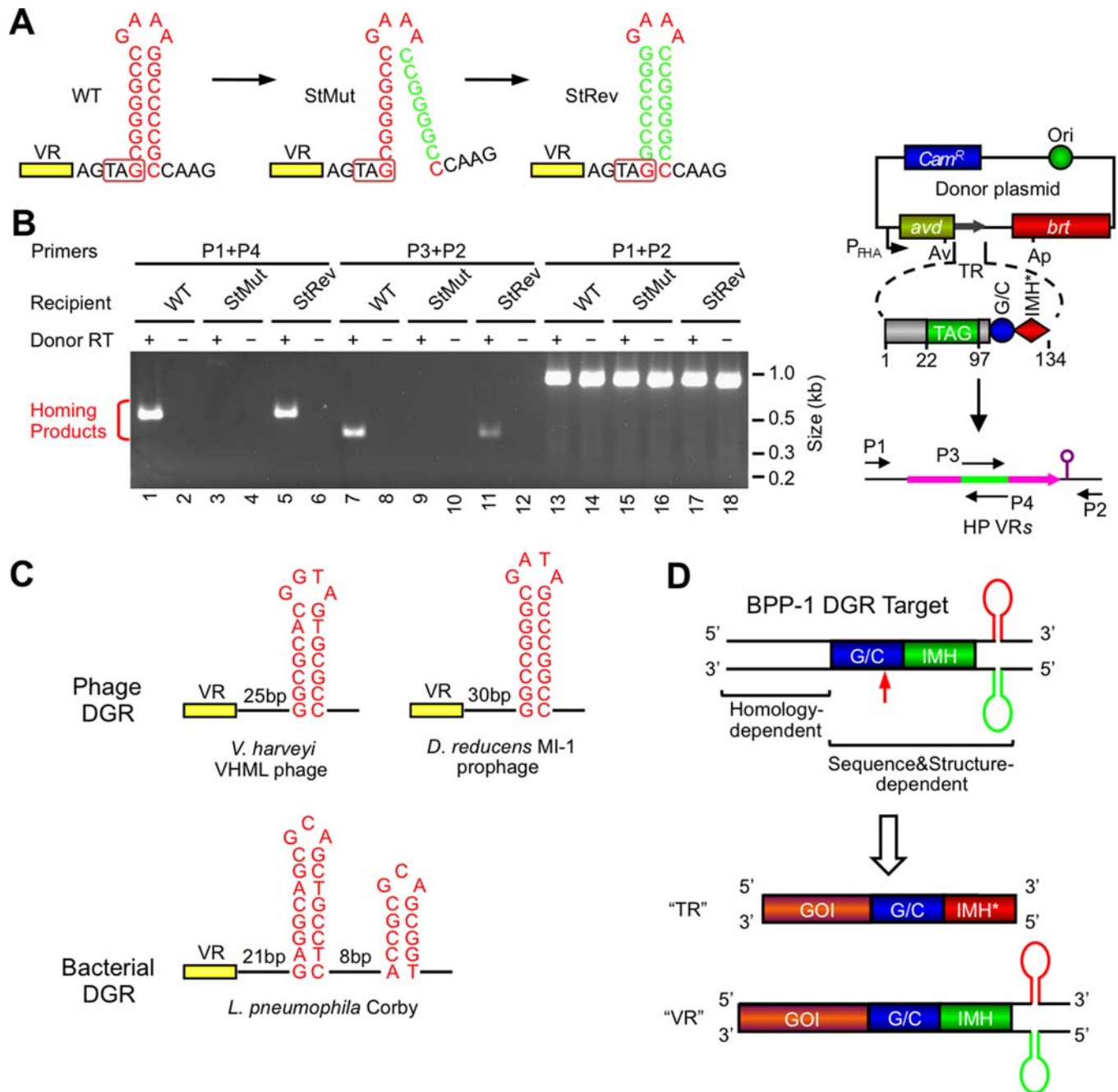


FIGURE 5 Role of a DNA secondary structure in DGR target recognition. **(A)** A DNA hairpin/cruciform structure downstream of the VR is required for BPP-1 DGR target recognition. The wild-type (WT) structure contains an 8-bp GC-rich stem and a 4-nt GAAA loop and is located 4 bp downstream of the VR. Mutating the 3' half of the stem (StMut) dramatically reduced DGR mutagenic homing **(B)** and phage tropism switching (not shown), while complementary changes to the 5' half of the stem (StRev) restored DGR activity in both assays. **(B)** PCR-based DGR homing assays with sequence-tagged TRs and VRs flanked by WT or mutant stem sequences. Shown on the right is a diagram of the PCR assay. Green represents the tag sequence transferred from the TR to the VR. P1 to P4 are primers annealing to the tag or flanking regions. **(C)** Similar DNA structures are found at analogous positions in a number of other phage (two shown) and bacterial (one shown) DGRs. The phage stems are GC rich and range from 7 to 10 bp, and loops have a conserved 4-nt sequence, G(A/G)NA. The *L. pneumophila* Corby DGR has a more complex tandem structure that is required for homing. **(D)** BPP-1 DGR target recognition at the 3' end is both sequence and structure dependent, requiring GC, IMH, and a hairpin/cruciform structure. Target recognition at the 5' end is homology mediated. By inserting a gene of interest (GOI) upstream of GC, IMH, and the DNA structure, the heterologous gene can be diversified by the BPP-1 DGR through appropriate engineering of the TR. (Adapted from 41.) [doi:10.1128/microbiolspec.MDNA3-0029-2014.f5](https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014.f5)

and phage tropism switching, while restoration of base pairing with heterologous complementary sequences restored DGR activity. This suggested that the DNA stem-loop structure, rather than its primary sequence, is important for target recognition. Further characterization of the BPP-1 structure showed that the length and G+C content of the stem modulate the efficiency of target recognition. The sequence of the 4-nt loop, however, appeared to be critical, as changes in either sequence or size dramatically reduced homing. Structure-specific nuclease digestion assays confirmed DNA structure formation *in vitro* in dsDNA, which required negative supercoiling. The position of the hairpin/cruciform, which is normally located 4 bp downstream of the VR, is also important. Moving it 4 bp in the 5' direction or 15 bp in the 3' direction significantly reduced DGR homing, although shorter insertions were tolerated. Marker transfer studies showed that extending the length between VR and the structured element did not affect the 3' marker co-conversion boundary, demonstrating that the cruciform influences the efficiency of homing, but not the site of cDNA integration. Although the exact function of the stem-loop structure is unknown, it is predicted to represent a binding site for host and/or DGR-encoded factors involved in cDNA priming and integration. Comparisons with other phage DGRs, as well as DGRs in bacterial chromosomes, suggested that similar structures are a conserved feature of target sequences (Fig. 5C). This was proved to be correct by Arambula *et al.* who demonstrated the required roles of analogous stem-loop sequences for homing by the *L. pneumophila* DGR (Figs 2 and 5C) (17).

As shown in Fig. 4, mutagenic homing is a “copy-and-replace” process that replaces parental VRs with TR-derived cDNAs that are mutagenized at specific sites (7). Target recognition at the 3' end is both sequence and structure dependent, while target recognition at the 5' end is dependent on short stretches of homology (Fig. 5D) (7, 41). A feature of the mechanism that seems key to its beneficial nature is that all *cis*-acting sequences and *trans*-acting factors required for additional rounds of diversification are precisely reconstituted during mutagenic homing. This allows repeated rounds of diversification and, presumably, the optimization of beneficial traits.

Guided by the TPRT model and the VR recognition rules described above (Fig. 5D), the BPP-1 DGR was engineered to target a kanamycin resistance gene (*aph3' Ia*) (41). A defective allele with a 3' truncation was placed upstream of the GC, IMH, and cruciform structure to form a recipient VR. A donor TR was then engineered to contain the missing *aph3' Ia* sequences along with a

short stretch of upstream homology. As predicted, the donor TR was able to repair the defective gene, conferring kanamycin resistance in an RT-dependent reaction. Thus, DGRs can be designed to target genes of interest, with potentially broad applications for protein engineering.

STRUCTURE AND FUNCTION OF THE Avd ACCESSORY PROTEIN

The BPP-1 DGR includes the *avd* locus, which encodes a positively charged 15-kDa protein that is essential for homing and is conserved in numerous bacterial and phage DGRs (6, 9, 10, 27). An X-ray crystal structure of BPP-1 Avd was determined by Alayyoubi *et al.* (Fig. 6A) to provide a guide for functional analysis (27). Avd forms a barrel-shaped homopentamer, which is positively charged on all surfaces, including an hourglass-shaped pore that runs through the center of the barrel and constricts to an ~8 Å diameter. Each Avd monomer forms a four-helix bundle with the α -helices running up and down in anti-parallel fashion.

A number of amino acid residues located on the surface or near the pore of the BPP-1 Avd pentamer were subjected to mutagenesis (27). Mutations at two conserved residues on the side of the pentamer, R79A and R83A, eliminated detectable homing, and mutations of two residues on the top of the pentamer, R36A and K37A, resulted in a partial loss of DGR activity. Defects in homing were not due to a loss of protein structure or stability, as identical CD spectra were obtained with wild-type and homing-defective Avd mutant proteins. Other mutations, including R19A (side), P35A (top), E43A (near the pore), and Q64A (bottom), had little effect on activity. These results suggest that positively charged residues on the side and top of the Avd pentamer have important functional roles.

As predicted by its positive charge, BPP-1 Avd was found to associate *in vitro* with single-stranded (ss)RNA, ssDNA, DNA:RNA complexes, and dsDNA, although the interactions were nonspecific (27). In contrast, Avd pentamers associated with purified Brt, which is also predicted to be positively charged, in a manner that was abolished by the R79A and R83A mutations (i.e., the ones that eliminated DGR homing), but not by other point mutations. These results suggested a correlation between Avd–Brt binding and DGR function. From these and other results, Alayyoubi *et al.* proposed that the pentameric nature of Avd may be involved in organizing a multivalent assembly consisting of Brt and nucleic acid components to somehow coordinate the multiple events required for homing (27).

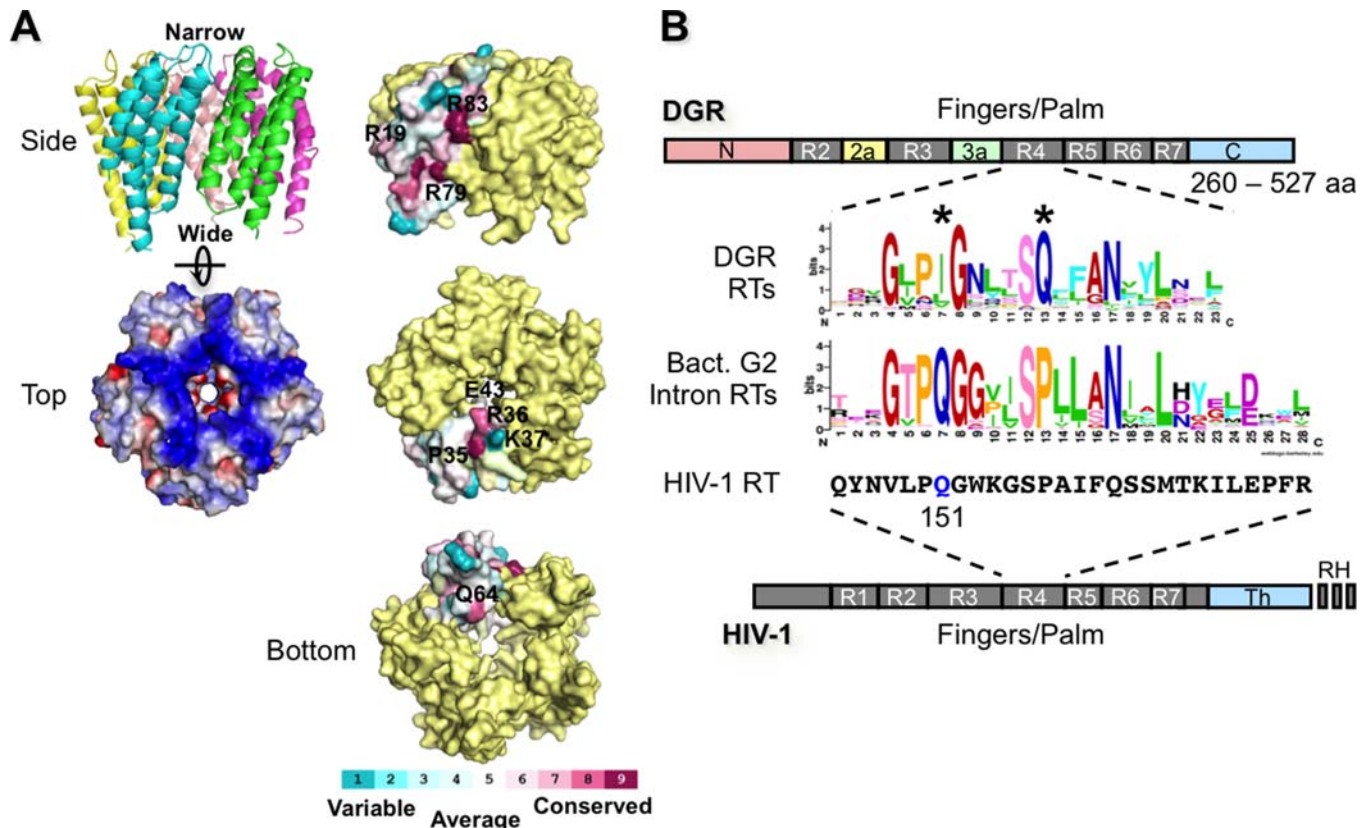


FIGURE 6 Avd and Brt. **(A)** Left: The BPP-1 Avd protein forms a homopentameric structure, with each monomer containing four helices running up and down (side view). The pentamer is highly positively charged (top view; blue, positively charged; red, negatively charged). Right: Amino acid residues on the side, top, and bottom of the Avd pentamer that were tested for Avd–Brt binding and/or DGR homing (27). **(B)** DGR RT domains and the sequence logo of its highly conserved domain R4. R1 to R7 are conserved sequence blocks found in the finger and palm domains of retroviral RTs, such as human immunodeficiency virus type 1 (HIV-1) RT (bottom). DGR RTs contain sequence insertions between R2 and R3 (R2a), and between R3 and R4 (R3a), as well as divergent N and C termini. They do not contain the thumb (Th) and RNase H (RH) domains that are found in HIV-1 RT. The domain R4 sequence logo of 155 DGR RTs was generated by Schillinger *et al.* using WebLogo (13, 42, 43). Comparison with the domain R4 sequence logo that we generated from 93 bacterial group II intron RTs (group II intron database: <http://webapps2.ucalgary.ca/~groupii/orf/orfalignment.html>; [44, 45, 46]), which are most closely related to DGR RTs, showed several characteristic differences, including the two highly conserved positions labeled with an asterisk (*). Also included for comparison is the corresponding amino acid sequence block of HIV-1 RT (strain BRU; GenBank accession no. K02013). The glutamine residue at position 151, which plays a role in nucleotide and template preference during reverse transcription, is highlighted in blue. (Adapted from 13.) doi:10.1128/microbiolspec.MDNA3-0029-2014.f6

DGR RTs AND ADENINE MUTAGENESIS

Adenine mutagenesis is a unique hallmark of DGRs, but its mechanism remains enigmatic. In the *td* intron-tagging experiment described above, no adenine-specific changes were observed in spliced RNA, suggesting that adenine mutagenesis does not involve RNA editing (7). However, adenine mutagenesis was observed in cDNA

products when VR1–99 was deleted to “trap” cDNA intermediates between the 3′ and 5′ integration steps, suggesting that adenine mutagenesis occurs during cDNA synthesis. These observations, coupled with the fact that DGR RTs are both highly conserved and unique, support the hypothesis that adenine mutagenesis is a property inherent to DGR RTs.

DGR RTs range from 260 to 527 amino acids, with an average length of ~380 residues (Fig. 6B) (13). They have divergent sequences at their N and C termini, with a conserved central core that includes common structural motifs found in most other RTs. They do not contain domains encoding RNase H activity, as found in retroviral RTs, or a DNA endonuclease activity as found in RTs of group II introns and LINEs. Within their shared sequences, DGR RTs have some intriguing features. The most prominent is located within the finger 4 region, which differs from those of group II intron RTs and human immunodeficiency virus type 1 (HIV-1) RT at several highly conserved positions (13). The finger 4 region of HIV-1 RT forms part of the nucleotide-binding pocket that positions incoming dNTPs and influences specificity and error-prone polymerization (13, 47). Mutations at Q151 (highlighted in blue in Fig. 6B) in HIV-1 RT, which corresponds to conserved isoleucine/leucine/valine residues in DGR RTs, change nucleotide and template preferences and confer resistance to inhibitory nucleoside analogs like azidothymidine (AZT) (48, 49, 50). It seems likely that unique features of this motif found in DGR RTs play a role in adenine mutagenesis.

At least two potential mechanisms could account for adenine mutagenesis. In the first, DGR RTs are hypothesized to carry out a variation of error-prone reverse transcription that inserts random dNTPs opposite adenines but rarely other residues, while synthesizing cDNA from TR-containing RNA templates. An alternative mechanism could be imagined in which DGR RTs have an increased propensity to incorporate dUTP, as opposed to dTTP when copying adenines. dUTP residues in cDNA products would then be recognized by host-encoded uracil DNA glycosylases and excised, leaving abasic sites. If these cDNA products are subsequently used as templates for second-strand cDNA synthesis, which could be catalyzed by a DGR RT or a host-encoded DNA polymerase, random nucleotides would be incorporated opposite these abasic sites.

Understanding the mechanism of adenine mutagenesis may have broad implications. During mutagenic homing by the BPP-1 DGR, about 30% of the TR adenines are converted to other nucleotides in progeny VRs (7). This error rate is far greater (~10,000 times higher) than that of HIV-1 RT, which is $\sim 1.4 \times 10^{-5}$ to 3.0×10^{-5} *in vivo* (51, 52). The low fidelity of HIV-1 RT enables the virus to generate vast numbers of sequence variants, enabling escape from immune surveillance and drug resistance. If DGR RTs are truly responsible for adenine mutagenesis, they are likely to be the most error-prone DNA-polymerizing enzymes yet discovered.

SUMMARY

DGRs are beneficial retroelements present in diverse bacteria and phage (6, 10, 11, 12, 13). Their apparent function is to accelerate the evolution of ligand-binding interactions. Structural, bioinformatic, and biochemical studies indicate that CLec folds are a conserved solution to display variable residues of DGR-diversified proteins (14, 20, 22). Unexpectedly, this structural fold is highly static in DGR-diversified proteins studied to date. Despite major differences in amino acid side chains, the backbones of BPP-1 Mtd variants are nearly identical to each other and their CLec folds superimpose with the diversified domain in *T. denticola* TvpA. This limited structural variability suggests that most DGR variable proteins will have relatively weak interactions with their respective ligands. For the BPP-1 phage, successful host recognition relies on multivalent interactions (i.e. avidity) between Mtd trimers on phage particles and arrays of receptors on the bacterial surface, which results in immense amplification of an otherwise weak monovalent ligand-receptor interaction. Avidity-driven binding is likely a conserved feature that contributes to the evolvability of DGR variable proteins (20).

A significant amount has been learned regarding the mechanism of DGR-mediated mutagenic homing (5, 6, 7, 17, 27, 41). Sequence diversification is mediated by a unique class of RTs associated with the conversion of adenine residues in an RNA intermediate into random nucleotides in a cDNA that ultimately replaces the variable region of a target gene with a diversified derivative. This process is independent of the host RecA-mediated homologous recombination machinery, similar to the mobility mechanism of group II introns (7). Current evidence is consistent with the hypothesis that mutagenic homing initiates through a TPRT mechanism, although other possibilities exist. BPP-1 DGR target recognition is both sequence and structure dependent at the 3' end, requiring GC and IMH sequence elements and a DNA cruciform structure, while target recognition at the 5' end is mediated by short stretches of homology (41). Based on these and other observations, the BPP-1 DGR was successfully engineered to target a heterologous gene.

Despite these advances, many questions remain unanswered. Hundreds of DGRs have been identified in diverse bacteria and phage, but their biological functions remain largely unknown (6, 10, 13, 17, 53). Understanding the functions of diverse DGRs is a major challenge. On the mechanistic side, the TPRT model is valuable for guiding experiments, but it is only one of several models that could explain DGR homing, and

the evidence supporting it is incomplete. In addition, it is unclear what specific role Avd plays in the homing reaction. The mechanism of adenine mutagenesis is of great interest as it represents a hallmark of DGR function. It has been hypothesized to be an intrinsic property of DGR-encoded RTs and, if true, understanding how it occurs may help us understand RT fidelity in general. Mutagenic homing is clearly a complex process that requires the participation of host factors, although none has been identified to date. Finally, the utility of DGRs for protein engineering has yet to be exploited. In addition to providing prodigious levels of diversity, mutagenic homing is a regenerative process that can operate through unlimited rounds to optimize protein functions. This may be particularly advantageous for directed protein evolution since desired traits can be selected and continuously evolved in iterative cycles, without the need for library construction or other interventions, and through a process that takes place entirely within bacterial cells.

ACKNOWLEDGMENTS

This work was supported by NIH grant (R01 AI069838) to J.F.M. and P.G. and University of Missouri startup funds to H.G. J.F.M. is a co-founder of AvidBiotics Corp. and a member of its Board of Directors and Scientific Advisory Board.

REFERENCES

- Vink C, Rudenko G, Seifert HS. 2012. Microbial antigenic variation mediated by homologous DNA recombination. *FEMS Microbiol Rev* 36:917–948.
- Agrawal A, Eastman QM, Schatz DG. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394:744–751.
- Kapitonov VV, Jurka J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol* 3:e181.
- Hencken CG, Li X, Craig NL. 2012. Functional characterization of an active Rag-like transposase. *Nat Struct Mol Biol* 19:834–836.
- Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, Maskell DJ, Simons RW, Cotter PA, Parkhill J, Miller JF. 2002. Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* 295:2091–2094.
- Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF. 2004. Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* 431:476–481.
- Guo H, Tse LV, Barbalat R, Sivaamnuaihorn S, Xu M, Doulatov S, Miller JF. 2008. Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Mol Cell* 31:813–823.
- Melvin JA, Scheller EV, Miller JF, Cotter PA. 2014. *Bordetella pertussis* pathogenesis: current and future challenges. *Nat Rev Microbiol* 12:274–288.
- Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, Baker S, Davis P, Simmonds M, Churcher C, Mungall K, Quail MA, Preston A, Harvill ET, Maskell DJ, Eiserling FA, Parkhill J, Miller JF. 2004. Genomic and genetic analysis of *Bordetella* bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol* 186:1503–1517.
- Medhekar B, Miller JF. 2007. Diversity-generating retroelements. *Curr Opin Microbiol* 10:388–395.
- Simon DM, Zimmerly S. 2008. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* 36:7219–7229.
- Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD. 2012. Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A* 109:3962–3966.
- Schillinger T, Lisfi M, Chi J, Cullum J, Zingler N. 2012. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics* 13:430.
- McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, Marti-Renom MA, Doulatov S, Narayanan E, Sali A, Miller JF, Ghosh P. 2005. The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat Struct Mol Biol* 12:886–892.
- Murphy K. 2012. *The Generation of Lymphocyte Antigen Receptors*, p. 171. In *Janeway's Immunobiology*, 8th ed. Garland Science, Taylor & Francis Group, LLC, London and New York.
- Abbas AK, Lichtman AH, Pillai S. 2012. *Lymphocyte Development and Antigen Receptor Gene Rearrangement*, p. 186. In *Cellular and Molecular Immunology*, 7th ed. Elsevier Saunders, Philadelphia, PA.
- Arambula D, Wong W, Medhekar BA, Guo H, Gingery M, Czornyj E, Liu M, Dey S, Ghosh P, Miller JF. 2013. Surface display of a massively variable lipoprotein by a *Legionella* diversity-generating retroelement. *Proc Natl Acad Sci U S A* 110:8212–8217.
- Stanley NR, Palmer T, Berks BC. 2000. The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in *Escherichia coli*. *J Biol Chem* 275:11591–11596.
- Narita S, Tokuda H. 2010. Sorting of bacterial lipoproteins to the outer membrane by the Lol system. *Methods Mol Biol* 619:117–129.
- Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, Ghosh P. 2008. Selective ligand recognition by a diversity-generating retroelement variable protein. *PLoS Biol* 6:e131.
- Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. 2010. Three-dimensional structure of tropism-switching *Bordetella* bacteriophage. *Proc Natl Acad Sci U S A* 107:4347–4352.
- Le Coq J, Ghosh P. 2011. Conservation of the C-type lectin fold for massive sequence variation in a *Treponema* diversity-generating retroelement. *Proc Natl Acad Sci U S A* 108:14649–14653.
- Boeke JD, Garfinkel DJ, Styles CA, Fink GR. 1985. Ty elements transpose through an RNA intermediate. *Cell* 40:491–500.
- Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* 87:917–927.
- Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* 94:451–462.
- Guo H, Karberg M, Long M, Jones JP, 3rd, Sullenger B, Lambowitz AM. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* 289:452–457.
- Alayyoubi M, Guo H, Dey S, Golnazarian T, Brooks GA, Rong A, Miller JF, Ghosh P. 2013. Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase. *Structure* 21:266–276.
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72:595–605.
- Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J* 21:5899–5910.
- Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* 82:545–554.

31. Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. 1995. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* 83:529–538.
32. Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3:a003616.
33. Kennell JC, Wang H, Lambowitz AM. 1994. The Mauriceville plasmid of *Neurospora* spp. uses novel mechanisms for initiating reverse transcription in vivo. *Mol Cell Biol* 14:3094–3107.
34. Chen B, Lambowitz AM. 1997. De novo and DNA primer-mediated initiation of cDNA synthesis by the Mauriceville retroplasmid reverse transcriptase involve recognition of a 3' CCA sequence. *J Mol Biol* 271:311–332.
35. George JA, Burke WD, Eickbush TH. 1996. Analysis of the 5' junctions of R2 insertions with the 28S gene: implications for non-LTR retrotransposition. *Genetics* 142:853–863.
36. Bibillo A, Eickbush TH. 2002. The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J Mol Biol* 316:459–473.
37. Bibillo A, Eickbush TH. 2004. End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J Biol Chem* 279:14945–14953.
38. Stage DE, Eickbush TH. 2009. Origin of nascent lineages and the mechanisms used to prime second-strand DNA synthesis in the R1 and R2 retrotransposons of *Drosophila*. *Genome Biol* 10:R49.
39. Zhuang F, Mastroianni M, White TB, Lambowitz AM. 2009. Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. *Proc Natl Acad Sci U S A* 106:18189–18194.
40. White TB, Lambowitz AM. 2012. The retrohoming of linear group II intron RNAs in *Drosophila melanogaster* occurs by both DNA ligase 4-dependent and -independent mechanisms. *PLoS Genet* 8:e1002534.
41. Guo H, Tse LV, Nieh AW, Czornyj E, Williams S, Oukil S, Liu VB, Miller JF. 2011. Target site recognition by a diversity-generating retroelement. *PLoS Genet* 7:e1002414.
42. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* 14:1188–1190.
43. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100.
44. Dai L, Toor N, Olson R, Keeping A, Zimmerly S. 2003. Database for mobile group II introns. *Nucleic Acids Res* 31:424–426.
45. Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S. 2008. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* 14:1704–1713.
46. Candales MA, Duong A, Hood KS, Li T, Neufeld RA, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. 2012. Database for bacterial group II introns. *Nucleic Acids Res* 40:D187–D190.
47. Huang H, Chopra R, Verdine GL, Harrison SC. 1998. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* 282:1669–1675.
48. Kaushik N, Talele TT, Pandey PK, Harris D, Yadav PN, Pandey VN. 2000. Role of glutamine 151 of human immunodeficiency virus type-1 reverse transcriptase in substrate selection as assessed by site-directed mutagenesis. *Biochemistry* 39:2912–2920.
49. Sarafianos SG, Hughes SH, Arnold E. 2004. Designing anti-AIDS drugs targeting the major mechanism of HIV-1 RT resistance to nucleoside analog drugs. *Int J Biochem Cell Biol* 36:1706–1715.
50. Boyer PL, Sarafianos SG, Clark PK, Arnold E, Hughes SH. 2006. Why do HIV-1 and HIV-2 use different pathways to develop AZT resistance? *PLoS Pathog* 2:e10.
51. Mansky LM, Temin HM. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 69:5087–5094.
52. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J Virol* 84:9864–9878.
53. Schillinger T, Zingler N. 2012. The low incidence of diversity-generating retroelements in sequenced genomes. *Mob Genet Elements* 2:287–291.