

UCLA

UCLA Electronic Theses and Dissertations

Title

Contribution of Cis-acting Elements and Trans-acting Host Factors in DGR-Mediated Mutagenic Homing

Permalink

<https://escholarship.org/uc/item/6mv1d6kj>

Author

Czornyj, Elizabeth

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Contribution of *Cis*-acting Elements and *Trans*-acting Host Factors in DGR-Mediated
Mutagenic Homing

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in Microbiology, Immunology and Molecular Genetics

by

Elizabeth Czornyj

2017

© Copyright by

Elizabeth Czornyj

2017

ABSTRACT OF THE DISSERTATION

Contribution of *Cis*-acting Elements and *Trans*-acting Host Factors in DGR-Mediated
Mutagenic Homing

by

Elizabeth Czornyj

Doctor of Philosophy in Microbiology, Immunology and Molecular Genetics

University of California, Los Angeles, 2017

Professor Jeffery Floyd Miller, Chair

Diversity-generating retroelements (DGRs) are family of retroelements that introduce nucleotide variability within defined protein-encoding DNA sequences. Sequence variation is site specific and occurs through a unique reverse transcriptase mediated process called mutagenic homing. DGRs were originally identified in the *Bordetella* phage BPP-1 and have since been identified in plasmids, bacteriophage and bacterial genomes. Moreover, DGRs were recently identified in Archaea and their viruses. Although DGRs are wide spread in nature and protein diversification has been demonstrated in both phage and bacterial systems, the precise mechanism of DGR

mutagenic homing remains to be elucidated. We have demonstrated that mutagenic homing requires specific nucleotide sequence and structural elements, including target site recognition sequences, which include a DNA stem-loop/cruciform structure. We recently demonstrated that in addition to base pairing interactions in the stem, the specific sequence and length of the 4nt loop are critical for DGR function. *In vitro* and *in vivo* analyses of the stem-loop structure indicate that the loop nucleotide composition has a major effect on stem-loop/cruciform formation and stability, and is thus critical for DGR function. In addition to influencing structure stability, we demonstrate that the orientation of the loop nucleotide sequence determines target site recognition during mutagenic homing. Stem-loops have been identified in most DGRs and our analysis of similar elements from disparate species indicates that these conserved elements are functionally interchangeable and fundamental to target site recognition. We propose that the stem-loop/cruciform structure serves as a recognition element for DNA processing events that culminate in cDNA synthesis, diversification, then integration. Since bioinformatic and functional analysis of DGRs reveals the lack of sequences predicted to encode enzymes involved in DNA or RNA processing events, we postulated host-encoded *trans*-acting factors play a pivotal role in mutagenic homing. To identify host-encoded factors that directly or indirectly influence the *Bordetella* phage BPP-1 DGR homing, a random transposon-insertion library was created. Individual transposon mutants were screened for insertions that had a significant effect on DGR homing as measured by a quantitative Km resistance assay. As an alternative approach to identify host factors, we performed targeted mutagenesis of candidate genes involved in DNA- and RNA-processing activities, including but not limited to ssDNA-specific exonuclease,

ATP-dependent DNA helicase, RNase H, RNase E, and other putative endoribonucleases. Mutants were screened for their ability to support mutagenic homing using phage tropism switching assays and we identified a subset of mutations in genes encoding DNA- and RNA-processing enzymes that decreased tropism switching. Taken together, our findings indicate that mutagenic homing involves both DGR-encoded and host-encoded factors that play a role in the diversification of target proteins, providing insight into a highly conserved mechanism for DNA editing.

The dissertation of Elizabeth Czornyj is approved.

Peter J. Bradley

James W. Gober

Elissa A. Hallem

Jeffery Floyd Miller, Committee chair

University of California, Los Angeles

2017

DEDICATION

I dedicate this work to my family: my parents, my brothers and sister, and John for their encouragement and support.

Table of Contents

ABSTRACT OF DISSERTATION	ii
ACKNOWLEDGEMENTS	ix
BIOGRAPHICAL SKETCH	x
CHAPTER 1. Introduction to Diversity Generating Retroelements	1
Diversity Generating Retroelements	3
Discovery of the <i>Bordetella</i> phage DGR	3
Genetic mechanism of DGR mutagenic homing	4
Target site recognition	5
Target protein scaffolds	6
Distribution of DGRs in Nature	8
Figure legends	15
References	22
CHAPTER 2. Target site recognition by a Diversity-Generating Retroelement (reprint)	25
CHAPTER 3. <i>Cis</i> -acting DNA Structural Elements Guide Targeted Mutagenesis by Diversity Generating Retroelements	
Abstract	59
Introduction	60
Results	63
Discussion	69

Figure legends	72
Materials and Methods	91
References	94

CHAPTER 4. Identification of Host-encoded Factors that Participate in DGR-mediated Mutagenic Homing

Abstract	99
Introduction	100
Results	103
Discussion	108
Figure legends	110
Materials and Methods	125
References	128

CHAPTER 5. Conclusion and Future Research

Conclusion	132
Future research	136
References	138

APPENDICES

APPENDIX A. Targeted diversity generation by intraterrestrial archaea and archaeal viruses (reprint)	141
APPENDIX B. A new topology of the HK97-like fold revealed in <i>Bordetella</i> bacteriophage by cryoEM at 3.5Å resolution (reprint)	150

Acknowledgements

I'm grateful to my mentor Dr. Jeff F. Miller for giving me the opportunity to be part of his lab and for all his support and guidance. I would also like to thank my committee for their advice during my graduate studies at UCLA.

I would like to thank the members of the JFM laboratory that have helped me throughout my stay in the lab. I'm especially thankful to Dr. Huatao Guo (University of Missouri, Columbia) for sharing his knowledge and insights. This work would not have been possible without him. My sincere gratitude to Dr. Baresi and Dr. Bishop for inspiring me to pursue a career in the field of Microbiology, and for their continued support and guidance.

The work presented in Chapter 2 is a reprint of Guo *et al.* "Target site recognition by a diversity-generating retroelement" with permission from Plos Genetics. The works included in the appendices are reprints of Blair *et al.* "Targeted diversity generation by intraterrestrial archaea and archaeal viruses" and of Zhang *et al.* "A new topology of the HK97-like fold revealed in *Bordetella* bacteriophage by cryoEM at 3.5 Å resolution" with permission of Nature Communications and eLife, respectively.

I'm immensely thankful to my entire family, especially my husband and best friend John, for encouraging me to achieve my goals and for offering their unconditional support over the years. I could not have done it without them.

Biographical Sketch

EDUCATION

- 9/2010 – Present University of California, Los Angeles, Los Angeles, CA
Ph.D. candidate, Microbiology, Immunology and
Molecular Genetics
- 9/2007 – 9/2010 California State University, Northridge, Northridge, CA
M.S., Biology
- 9/2004 – 10/2007 California State University, Northridge, Northridge, CA
B.S., Biology
- 2002 – 2004 Los Angeles Community College, Los Angeles, CA
A.S., Chemistry

RESEARCH EXPERIENCE

- 2011– Present Jeff F. Miller, MIMG Department, University of California,
Los Angeles, CA
Graduate Student Researcher
- 9/2007– 9/2010 Larry Baresi, Department of Biology, California State
University Northridge, Northridge, CA
Graduate Student Researcher
- 2004 – 2007 Larry Baresi, Department of Biology, California State
University Northridge, Northridge, CA
Undergraduate Student Researcher

PUBLICATIONS

1. Toso DB, Javed MM, **Czornyj E**, Gunsalus RP, Zhou ZH. 2016. Discovery and characterization of iron-sulfide and polyphosphate bodies co-existing in *Archaeoglobus fulgidus* cells. *Archaea* doi:10.1155/2016/4706532.
2. Blair GP, Bagby SC, **Czornyj E**, Arambula D, Handa S, Sczyrba A, Ghosh P, Miller J.F, Valentine DL. 2105. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nature Communications* **6**: 6585.
3. Zhang X, Guo H, Jin L, **Czornyj E**, Hodes A, Hui WH, Zhou ZH. 2013. A new topology of the HK97-like fold revealed in *Bordetella* bacteriophage by cryoEM at 3.5 Å resolution. *eLife* **2**: e01299.

4. Arambula D, Wong W, Medhekar B, 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* **110**: 8212-8217.

5. 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 Genetics* **7**: 12.

POSTERS

- May 2014. Identification of Host-encoded Factors Involved in the regulation of DGR-mediated Mutagenic Homing. American Society for Microbiology (ASM) general meeting, Boston, MA.
- June 2012. Diversity-Generating Retroelements: Target Site Recognition and Distribution in Nature. American Society for Microbiology (ASM) general meeting, San Francisco, CA.
- September 2011. Diversity-Generating Retroelements promote accelerated evolution of target genes in bacteria and phage, and have broad implications for protein engineering. UCLA MIMG 2011 retreat, Los Angeles, CA.
- May 2011. Protein and mRNA expression during Methanophage G infection of *Methanobrevibacter* strain G. American Society for Microbiology (ASM) general meeting, New Orleans, LA.
- November 2007. Isolation of methanogen inhibitors from bovine rumen. Southern California Branch of American Society for Microbiology (SCASM) annual meeting, San Diego, CA.

HONORS AND AWARDS

- University Fellowship, University of California, Los Angeles
- Eugene V. Cota-Robles Fellowship, University of California, Los Angeles
- Academic achievement award, California State University, Northridge

CHAPTER 1. Introduction to Diversity Generating Retroelements

Retroelements (REs) are mobile genetic elements that are broadly distributed in prokaryotic and eukaryotic organisms [1]. REs are a class of transposable elements that encode a reverse transcriptase which functions during replicative transposition through an RNA intermediate and disseminate within the host genome by a “copy and paste” mechanism [1]. REs are divided into two major groups based on the presence or absence of long terminal repeats (LTRs). LTR-containing retroelements comprise ~8% of the human genome, and include LTR retrotransposons, tyrosine recombinase retrotransposons, and endogenous retroviruses [1]. The second group is known as non-LTR retroelements and includes long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and processed pseudogenes. Non-LTR retroelements are proposed to have derived from group II introns [2, 3] and both systems appear to have a similar mechanism of action [1-4].

Group II introns are found in chloroplast and mitochondrial genomes of fungi and plants and are common in both gram-negative and gram-positive bacteria [2]. Group II introns consist of a catalytically active intron RNA and an intron-encoded protein (IEP), which has reverse transcriptase activity [2, 3]. Mobility of group II introns occurs by a target DNA-primed reverse transcription (TRTP) mechanism in which the intron RNA reverse splices directly into a DNA target site and is then reverse transcribed into cDNA by the IEP [2, 3].

Diversity-generating retroelements (DGRs) are a recently discovered class of retroelements that appear to have been derived from group II introns [1, 5].

Diversity Generating Retroelements

DGRs are a family of retroelements capable of generating variability within protein-encoding DNA sequences [6, 7]. DGRs are unique among retroelements due to their potential to confer selective advantages by dramatically accelerating the evolution of adaptive traits through a distinct reverse transcriptase mediated process called mutagenic homing [6-8]. DGRs are composed of highly conserved components such as a dedicated reverse transcriptase (RT), an accessory gene (*avd* or *HRDC*), a template repeat (TR), a variable repeat (VR) and a unique target gene that is subject to diversification. DGRs function through a template-dependent, reverse transcriptase (RT)-mediated process which introduces nucleotide substitutions at defined sites within protein-encoding DNA sequences [6, 7]. Mechanistic studies of the prototype DGR, BPP-1, have provided insights into the mechanism by which diversification of target proteins occur and will be discussed below.

Discovery of the *Bordetella* phage DGR

The prototypic DGR was discovered in the bacteriophage BPP-1, which infects the mammalian respiratory pathogen *Bordetella bronchiseptica* (*Bb*) [6]. The infectious cycle of *Bb* is regulated by the two-component system, BvgAS, which alternates between active (Bvg⁺) and inactive states (Bvg⁻) to modulate expression of cell surface molecules in order to mediate respiratory infection or *ex vivo* survival, respectively [9]. The BPP-1 (Bvg plus-tropic phage 1) preferentially infects Bvg⁺ cells through the receptor pertactin (Prn), an outer membrane transporter protein only expressed in the Bvg⁺ phase [6]. However, at a frequency of $\sim 10^{-6}$, BPP-1 forms normal size plaques on

Bvg⁻ phase *Bordetella* [6]. Since plaque formation requires multiple rounds of phage infection and multiplication, the observation that BPP-1 was capable of infecting Bvg⁻ phase cells suggested that a tropism switch had occurred [6]. In fact, two types of tropic variants were identified. The first was designated BMP (Bvg minus-tropic phage) and preferentially infects Bvg⁻ phase cells. The second was designated BIP (Bvg indiscriminant phage) and infects both Bvg⁺ and Bvg⁻ phase cells with nearly equal efficiencies [6]. The observation that BPP-1 could generate tropic variants that recognized distinct cell surface molecules led to the discovery that tropism switching, or altered ligand for infection, is mediated by the phage-encoded DGR.

The BPP-1 DGR facilitates tropism switching by introducing nucleotide substitutions in a gene that specifies a host cell-binding protein, Mtd (major tropism determinant), positioned at the distal tips of phage tail fibers that facilitate binding to cell surface molecules [6, 10, 11]. Nucleotide substitutions in Mtd results in tail fibers with distinct receptor specificity, allowing BPP-1 adaptation to the dynamic changes in cell surface molecules that occur during the host infectious cycle [10, 12].

Genetic mechanism of DGR mutagenic homing

Using the BPP-1 DGR as an archetype, a mechanism for DGR mutagenic homing has been proposed (Figure 1). BPP-1 tropism switching is mediated by nucleotide substitutions in a 134 base pair (bp) variable repeat (VR) located at the 3' end of the target gene, *mtd* (Figure 1). Mtd is a trimeric tail fiber protein that mediates phage binding to *Bordetella* surface molecules and changes in its coding sequence result in new ligand specificities [10]. Sites of nucleotide substitutions in VR correspond to

adenine nucleotides in a homologous template repeat (TR), which is invariant and serves as a template to derive an RNA intermediate. During reverse transcription of the TR-derived RNA intermediate by the bRT (*Bordetella* reverse transcriptase), adenine residues are replaced with random nucleotides and the synthesized cDNA displaces the parental VR in a process termed mutagenic homing [13]. Mutagenic homing is proposed to occur through a unique target-primed reverse transcription (TPRT) mechanism similar to that of group II introns in bacteria [13]. We postulate that homing initiates with either a single strand nick or a double-stranded break in the IMH (initiation of mutagenic homing), which is a DNA region located at the 3' end of VR. The resulting 3' hydroxyl serves as a primer to reverse-transcribe the TR-derived RNA intermediate [13]. This process generates a cDNA product that replaces the parental VR with 3' integration occurring at the nick/double-stranded break and integration at the 5' end of VR depending on TR-VR homology. Although not definitive, we predict replacement of the parental VR may occur via template switching or strand displacement. The replication of the resulting cDNA-VR heteroduplex generates progeny genomes with mutagenized VRs. Although the proposed model is consistent with present data, the basis of adenine mutagenesis and the mechanism of cDNA integration as well as the function of IMH are unknown.

Target site recognition

Recent studies identified the *cis*-acting sequences required for a gene to be recognized as a target [14]. In order for a gene to be diversified, the gene must contain short stretches of homology with its cognate TR and must be recognized by the DGR

machinery [13]. In the BBP-1 DGR, these requirements are provided by the IMH element (Figure 3) [7, 13]. The BPP-1 IMH is composed of a 14 bp G/C stretch (G/C₁₄) which is identical to the corresponding segment of TR and a 21 bp segment which defines the extent of homology with TR [7]. Moreover, the site at which the TR-derived sequence information is incorporated into VR was mapped to the G/C₁₄ element (Figure 2) [13]. A third element was recently identified. This 24 bp sequence contains an inverted repeat that under physiological levels of negative supercoiling forms a stem-loop or double-stranded cruciform structure with an 8bp stem and a 4nt loop (Figure 2). We have recently demonstrated that stem-loop/cruciform formation is essential for efficient target site recognition during mutagenic homing [14].

Target protein scaffolds

The BPP-1 DGR system has been extensively characterized and the target protein has been both functionally and structurally analyzed. To understand the structural biology of variable proteins, McMahon *et al.* determined the crystal structures of five Mtd variants (Figure 3) [10]. These structures show that Mtd variants were nearly identical in overall structure and formed tetrahedral-shaped homotrimers, which are positioned at the distal ends of phage tail fibers [10,11]. Moreover, the VR-encoded variable amino acids are contained within a C-type lectin-like (CLec) domain located at the C-terminus of each Mtd monomer. All variable residues in Mtd are solvent exposed and they form a discrete ligand-binding surface on the bottom face of each monomer (Figure 4A and B) [10]. Interestingly, comparative analysis of Mtd structures revealed that although each of the five Mtd tropic variants displayed different VR surface residues, the conformations of the

VR-encoded regions were identical (Figure 4C) [10]. The BPP-1 TR contains 23 adenine residues subject to mutagenesis, which can theoretically generate about 10^{14} (4^{23}) different VR sequences [6].

Although target proteins vary in size and predicted function, VR regions are nearly always located at the C-termini, and diversification of target proteins often results in amino acid substitutions or variable residues that are displayed within C-Lec scaffolds [10, 11]. This was demonstrated in the DGR variable protein from *Treponema denticola*, TvpA, which is predicted to be a lipoprotein localized on the outer surface of the gram-negative spirochete. The variable domain in TvpA is organized in a C-Lec fold and the adenines in TR correspond to variable residues in VR that form the ligand-binding pocket of the target protein [15]. This indicated that nucleotide diversity is constrained within conserved domains of target proteins, resulting in a balance of protein diversity and scaffold stability.

Furthermore, analysis of phage-encoded DGRs has recently identified a subset of hypervariable regions located within immunoglobulin (Ig) domains similar to those found in antibodies and T cell receptors [16]. In contrast to CLec folds, which are usually located near the 3' end of target genes, these Ig folds were located in the middle of the VR-containing ORFs. This suggests that DGRs have evolved to use different protein scaffolds that are designed to accommodate massive amino acid variability.

Distribution of DGRs in Nature

Although first identified in the *Bordetella* phage BPP-1 [6], DGRs are widely distributed in nature and have been identified in plasmids, bacteriophage and bacterial genomes (Figure 5) [5, 7, 8, 16, 17]. DGRs present in sequence databases have been identified using custom made algorithms, including DiGReF which identifies putative RTs. Following identification of RTs, flanking nucleotide sequences are analyzed to identify VR/TR pairs that exclusively differ at positions corresponding to TR-adenines. To date, over 1300 DGRs have been identified in organisms that occupy widely diverse ecological niches and are found in organisms that have planktonic, symbiotic and pathogenic lifestyles. These include human pathogens such as *Legionella pneumophila* and *Treponema denticola* as well as human commensals such as *Bacteroides spp.* Moreover, DGRs were recently identified in Archaea and their viruses [18]. The breadth of organisms that contain DGRs, along with their divergent lifestyles, indicate that they represent a conserved prokaryotic system for targeted protein evolution. A brief overview of DGRs encoded by phage, and bacterial and archaeal genomes is discussed below.

Phage-encoded DGRs

Since DGRs were first discovered in the *Bordetella* phage BPP1, homologous elements have been identified in numerous phage [7, 8, 16, 18]. Phage-encoded DGRs are found near tail fiber genes and target genes are predicted or known to encode phage tail fiber proteins [11, 27].

A recent study by Minot *et al.* has identified a vast number of DGRs associated with phage that occupy the human intestinal gut [16]. Analysis of the human gut virome of healthy individuals identified DGRs in phage populations present in 11 out of 12 subjects studied. Metagenomic analysis of phage populations from stool samples identified 51 variable regions, 36 of which contained a VR/TR pair. Of these 36 variable regions, 29 contained a TR/VR pair near an RT gene, which differs at sites corresponding to TR adenines, a hallmark of DGR functionality. A subset of these 29 hypervariable regions were located at the 3' end of the target genes and were predicted to be contained within a C-Lec fold, similar to the Mtd of BPP-1. Interestingly, six hypervariable regions were found within a Ig β -sandwich domain located in the middle of the target ORFs as opposed to the 3' ends [16]. To our knowledge, this is the first described DGR system known to target Ig folds for sequence variation.

Furthermore, the first archaeal DGR system was recently identified in a novel archaeal virus called ANMV-1 [18]. The ANMV-1 DGR encodes conserved components required for mutagenic homing, including an RT, a VR/TR pair that vary from each other at positions corresponding to adenines, a putative accessory protein (Avd) and a short inverted repeat that has the potential to form a stem-loop/cruciform structure [18]. As in other DGR target proteins, variable residues of the ANMV-1 DGR target protein, AdtA, are contained within a CLec fold which suggests a ligand-binding role for AdtA.

Moreover, the *adtA* gene is located in close proximity to phage tail fiber genes and AdtA might function as a receptor binding phage tail fiber protein.

Bacterial-encoded DGRs

DGRs are ubiquitous in bacterial genomes and they have been identified in members of the human microbiome (e.g. *Bacteroides fragilis*), numerous pathogens of mammals or plants (e.g. *L. pneumophila*, *T. denticola*, *Ralstonia* spp.) and organisms of environmental importance such as Cyanobacteria, which produce 50% of the earth's oxygen (Figure 4) [6-8]. Furthermore, DGRs appear to be enriched in three phyla of bacteria: Bacteroidetes, Firmicutes and Proteobacteria [5]. These elements are located within prophage, plasmids, and conjugative transposons suggesting that DGRs use different modes for dispersal and can be horizontally transferred.

DGRs in *Legionella* species

L. pneumophila is a facultative intracellular parasite of protozoa and an opportunistic human pathogen that causes both community- and hospital-acquired pneumonia [19, 20]. DGRs have been identified in a subset of *L. pneumophila* clinical isolates that includes *L. pneumophila* Corby, *L. pneumophila* D5572 and *L. pneumophila* D5591 [21]. Characterization of the *L. pneumophila* DGR in strain Corby has shown that the DGR cassette is found in a genomic island located within an integrative and conjugative element (ICE) [21]. The *L. pneumophila* Corby DGR encodes all *cis*- and *trans*-acting elements required for mutagenic homing, including RT, Avd, VR, TR, IMH and IMH* as well as stem-loop structures downstream of VR [21]. Moreover, adenine-specific mutagenic homing and diversification of its target protein, LdtA, has been demonstrated [21]. LdtA is a lipoprotein that contains a twin-arginine translocation (TAT) and a lipobox motif at its N-terminus. The VR is located at the C-terminus of LdtA and variable

residues are accommodated within a CLec fold. Furthermore, LtdA is a TAT-secreted protein that is localized to the outer leaflet of the outer membrane with its diversified C-terminal sequence surface-exposed [21]. A remarkable feature of the *L. pneumophila* Corby DGR is that its TR contains 43 adenines, corresponding to a DNA diversity in VR of $\sim 10^{26}$ sequences which vastly exceeds the capabilities of the vertebrate adaptive immune system.

DGRs in *Treponema* species

Genomic and metagenomic analyses have identified multiple DGRs in *T. denticola* [15, 22]. *T. denticola* is a human oral pathogen that is associated with periodontal disease [23]. More recently, analysis of *T. denticola* strains obtained from the NCBI and the Human Microbiome Project (HMP) datasets identified complete DGRs in 9 of the 17 assembled *T. denticola* genomes [22]. These elements encode characteristic features of a DGR, which include RT, TR and VR repeats that differ at positions corresponding to TR-adenines. Some of the DGRs encode seven target genes with VRs that can be potentially diversified from the same TR. Moreover, all identified target proteins have a CLec domain and appear to be similar to TvpA (*Treponema* variable protein A) [15, 22]. The TvpA target proteins contain lipoprotein signal sequences that are predicted to target TvpA to the outer surface of the bacterium [15, 22].

DGRs in *Bacteroides* species

Bacteroides are members of the human gut microbiome and DGRs have been shown to be particularly abundant in the Bacteroidetes phylum [5]. DGRs in *Bacteroides* species

have been found in prophage, plasmids and conjugative transposon elements that might be used for DGR transmission between species [5]. Although *Bacteroides* DGRs appear to be widely distributed in the human gut microbiome, little is known about their function. While analysis of the *B. fragilis* 638R DGR is ongoing, bioinformatic analysis indicates that this element encodes conserved features required for mutagenic homing and is located within an integrative and conjugative element (ICE) [Yanling Wang personal communication]. Consistent with features of DGR target proteins, the VR is located at the C-terminus of the target protein, BfdT (*B. fragilis* DGR target protein), and the variable residues are contained within a CLec fold. The BfdT contains a N-terminal lipobox motif and is proposed to be a surface exposed lipoprotein similar to the LdtA target protein of the *L. pneumophila* DGR. *Bacteroides* DGRs encode conserved components (Figure 5), including an RT, TR and VR sequences as well as an accessory protein (Avd or Ch). Moreover, *cis*-acting elements have been identified near the 3' end of the target gene and include an inverted repeat(s) that forms a stem-loop/cruciform structure. Moreover, our recent analysis indicates that the *Bacteroides ovatus* stem-loop is capable of supporting mutagenic homing in BPP-1, which suggests that stem-loops are conserved structures that serve as target recognition elements across species (see chapter 3).

Archaeal-encoded DGRs

DGRs have been recently identified in Archaea and their viruses [18]. Analysis of genomic datasets identified multiple DGRs in the genomes of uncultivated subterranean *Nanoarchaeota* [18]. These DGR systems contain *cis*- and *trans*- acting elements

required for mutagenic homing. Furthermore, a recent inquiry into metagenome datasets uncovered hundreds of additional DGRs in archaeal and bacterial genomes, and were found as being prevalent within the archaeal superphylum, DPANN (Pacearchaeota and Woesearchaeota) and the bacterial candidate phyla radiation (CPR) (Blair Paul personal communication).

Although DGRs are wide spread in nature and protein diversification has been demonstrated in both phage and bacterial systems, the precise mechanism of mutagenic homing has not been determined. As mentioned above, *cis*-acting elements are necessary for a gene to be recognized as a target by the DGR machinery. These include an inverted repeat that forms a stem-loop or cruciform structure. The sequence requirements for stem-loop formation and the contribution of this structural element in target site recognition will be described in Chapter 2.

While the formation of a stem-loop/cruciform structure is necessary for efficient nucleotide sequence recognition of target genes, the precise role stem-loops play in mutagenic homing has not been determined. Using the model system BPP-1, we investigated how the formation, structure and stability of DGR stem-loops contribute to target site recognition and our findings will be discussed in Chapter 3 (currently in preparation for publication).

With the notable exception of a unique family of reverse transcriptases, both bioinformatic and functional analysis of DGRs reveals the lack of sequences predicted to encode enzymes that participate in DNA- or RNA-processing events that lead to

cDNA synthesis, diversification and then integration. In Chapter 4, we describe a *Bordetella bronchiseptica* genetic screen to identify host-encoded factors that play a role in mutagenic homing in the *Bordetella* phage BPP-1 DGR.

Figure legends:

Figure 1. *Bordetella* phage BPP-1 DGR mutagenic homing model.

In the current model, BPP-1 phage DGR mutagenic homing occurs through a target-primed reverse transcription mechanism [13]. DGRs diversify DNA sequences through a process called mutagenic homing, which introduces nucleotide substitutions into the VR (green arrow) of the target gene, *mtd* (green). Mutagenic homing initiates with either a single strand nick or a double-stranded break in the IMH (pink), a DNA region located at the 3' end of VR, and the resulting 3' hydroxyl serves as a primer to reverse-transcribe the TR-derived RNA intermediate (blue). The TR-RNA provides a template for DGR-RT (red box) dependent cDNA synthesis. During reverse transcription, TR-adenines (A, red) are randomly changed to any of the four nucleotides (N, red), which result in a mutagenized cDNA that displaces the parental VR at the 3' end of Mtd. The Mtd is located at the distal tail fibers of the phage (green circles), and VR diversification results in Mtd variants (red circles) that recognize new host-cell surface molecules as ligands for infection. Avd (aqua box) encodes an accessory protein that is proposed to interact with RT and nucleic acids [24].

Figure 2. Components of the IMH element in BPP-1 DGR.

Components of the IMH (Initiation of Mutagenic Homing) are depicted in the expanded view and include a 14bp G/C-rich region, a 21bp sequence and an inverted repeat that forms a stem-loop structure composed of a 8bp stem and a 4nt loop. Only a single strand is represented. The red arrow indicates the putative cDNA initiation site. Extent of homology with TR is shown. Box, Mtd stop codon.

Figure 3. Structural features of Mtd [10].

a) Mtd trimer with 3 variable regions (VR) located on the bottom face. b) Ligand-binding site of an Mtd variant with VR residues in red. c) Superimposed VRs from five Mtd variants. Main chain conformation and variable side chains of superimposed VR-encoded sequences are shown.

Figure 4. Phylogenetic Distribution of DGRs in Bacteria.

The 16s rRNA-based tree shows phyla in which DGR-containing bacteria were identified (gray boxes). The genera and species containing DGRs are indicated and those containing chromosomal DGRs are shown in red. Phage and plasmid DGRs are shown in blue.

Figure 5. DGR cassette architectures in *Bacteroides* spp.

Schematics of *Bacteroides* DGRs. a) Putative *cis*- and *trans*- acting features of the *B. ovatus* DGR. The target protein (TP) is shown as a purple box, and the RT and the accessory variability determinant (avd) are shown as red and aqua boxes, respectively. Black arrows indicate VR and TR. Located at the 3' end of VR is an inverted repeat that has the potential to form a stem-loop/cruciform structure. b) Schematic of the *B. fragilis* 638R DGR cassette. The target protein, bftA, a VR/TR pair (black arrows) and RT are shown. The DGR encodes two accessory proteins (CH and AVD) and two inverted repeats that have the potential to form stem-loop or cruciform structures.

Figure 1.

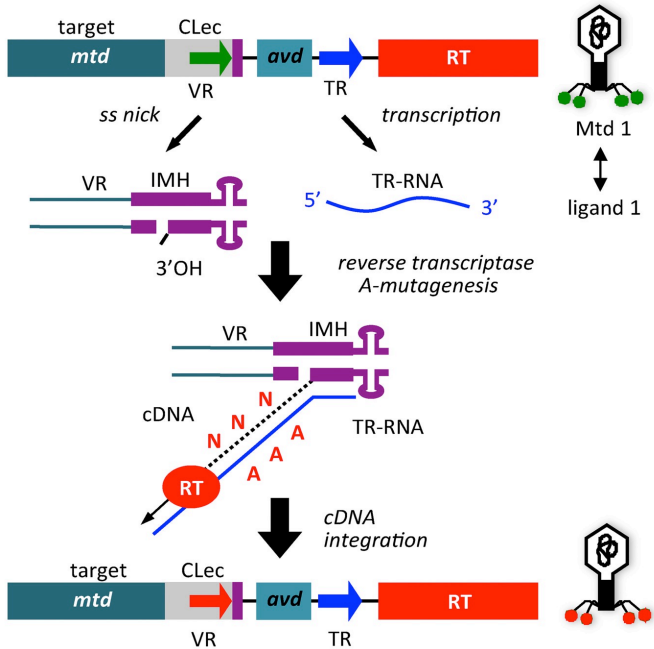


Figure 2.

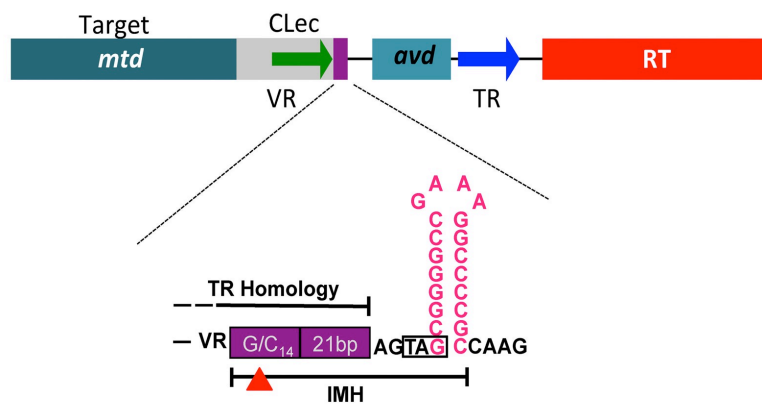


Figure 3.

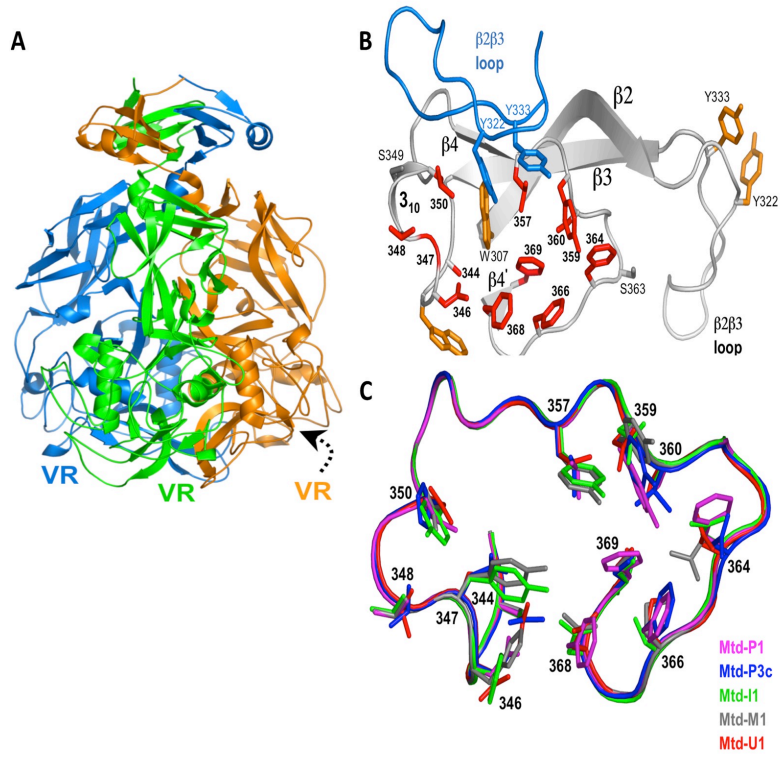


Figure 4.

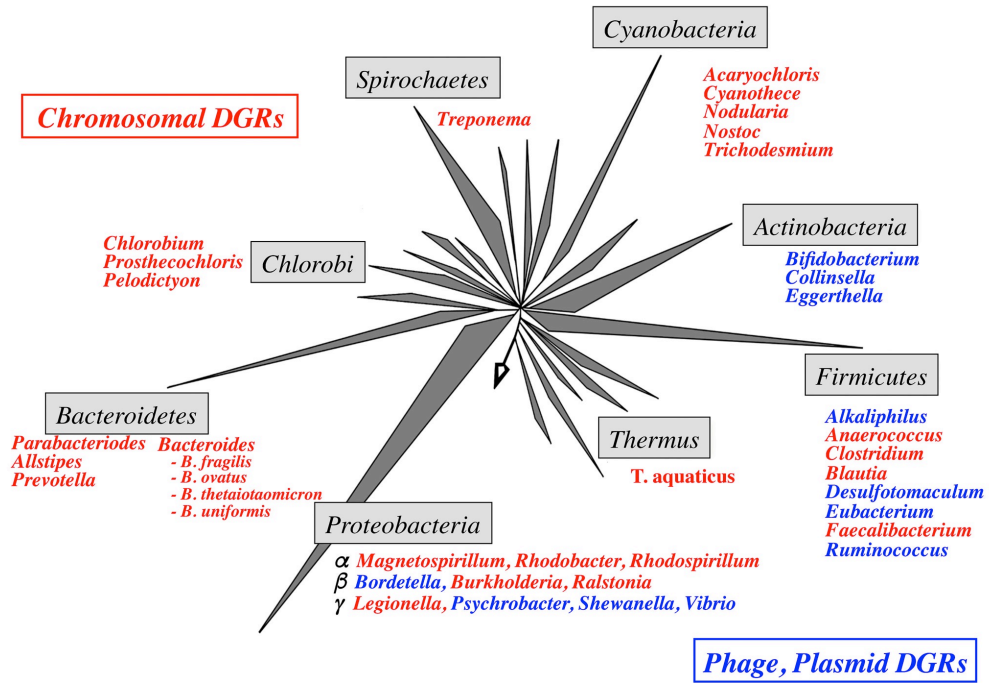
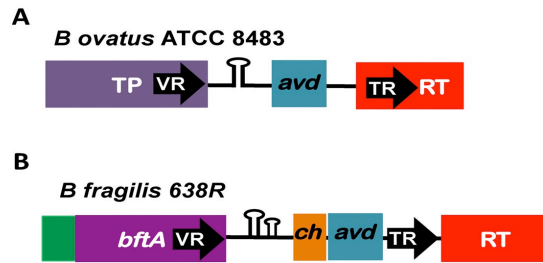


Figure 5.



REFERENCES

1. Gogvadze E, Buzdin A. 2009. Retroelements and their impact on genome evolution and functioning. *Cell. Mol. Life Sci.* **66**: 3727-3742.
2. Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. *Ann Rev Genet* **38**: 1-35.
3. Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol.* **3**(8): a003616.
4. Toro N, Jimenez-Zurdo JI, Garcia-Rodriguez FM. 2007. Bacterial group II introns: not just splicing. *FEMS Microbiol Rev* **31**: 342-358.
5. 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.
6. 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.
7. Doulatov S, Hodes A, Dai L, Mandan 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.
8. Medhekar B, Miller JF. 2007. Diversity-generating retroelements. *Current Opinion in Microbiology* **103**: 88-395.

9. Williams CL, Boucher PE, Stibitz S, Cotter PA. 2005. BvgA functions as both an activator and a repressor to control Bvg phase expression of *bipA* in *Bordetella pertussis*. *Mol Microbiol* **56**: 175-188.
10. 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.
11. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. 2010. Three-dimensional structure of tropism-switching *Bordetella* bacteriophage. *Proc Natl Acad U S A* **107**(9): 4347-4352.
12. 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**: 131. PMID: PMC2408619.
13. 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.
14. Guo H, Tse LV, Nieh A, Czornyj E, Williams S, Oukil S, Lieu V, Martin D, Miller JF. 2011. Sequence Requirements for Target-Site Recognition by a Diversity-Generating Retroelement. *PLoS Genetics* **7**: 12.
15. Le Coq J, Ghosh P. 2011. Conservation of the C-type lectin fold for massive sequence variation in a *Treponema* diversity generating retroelement. *PNAS* **108**(35): 14649-14653.

16. 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.
17. Simon DM, Zimmerly S. 2008. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* **36**:7219-7229.
18. Paul BG, Bagby S, Czornyj E, Arambula D, Handa S, Sczyrba A, Ghosh P, Miller JF, Valentine DL. 2015. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nat. Commun* **6**:6585.
19. Newton HJ, Ang DK, van Driel IR, Hartland EL. 2010. Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clin Microbiol Rev.* **23**: 274-298.
20. Carratalà J, Garcia-Vidal C. 2010. An update on Legionella. *Curr Opin Infect Dis.* **23**: 152-7. PMID: 20051846.
21. 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 U S A* **110**(20): 8212-8217.
22. Nimkulrat S, Lee H, Doak TG, Ye Y. 2016. Genomic and Metagenomic analysis of Diversity-Generating Retroelements Associated with *Treponema denticola*. *Front Microbiol* **7**: 852.
23. Loesche WJ, Grossman NS. 2001. Periodontal disease as a specific, albeit chronic, infection: Diagnosis and treatment. *Clin Microbiol Rev* **14**: 727-752.
24. 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**(2): 266-167.

CHAPTER 2. Target Site Recognition by a Diversity-Generating Retroelement

Target Site Recognition by a Diversity-Generating Retroelement

Huatao Guo¹, Longping V. Tse¹, Angela W. Nieh¹, Elizabeth Czornyj¹, Steven Williams², Sabrina Oukil¹, Vincent B. Liu¹, Jeff F. Miller^{1,3*}

1 Department of Microbiology, Immunology, and Molecular Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America, **2** AvidBiotics Corporation, South San Francisco, California, United States of America, **3** The Molecular Biology Institute, University of California Los Angeles, Los Angeles, California, United States of America

Abstract

Diversity-generating retroelements (DGRs) are *in vivo* sequence diversification machines that are widely distributed in bacterial, phage, and plasmid genomes. They function to introduce vast amounts of targeted diversity into protein-encoding DNA sequences via mutagenic homing. Adenine residues are converted to random nucleotides in a retrotransposition process from a donor template repeat (TR) to a recipient variable repeat (VR). Using the *Bordetella* bacteriophage BPP-1 element as a prototype, we have characterized requirements for DGR target site function. Although sequences upstream of VR are dispensable, a 24 bp sequence immediately downstream of VR, which contains short inverted repeats, is required for efficient retrohoming. The inverted repeats form a hairpin or cruciform structure and mutational analysis demonstrated that, while the structure of the stem is important, its sequence can vary. In contrast, the loop has a sequence-dependent function. Structure-specific nuclease digestion confirmed the existence of a DNA hairpin/cruciform, and marker coconversion assays demonstrated that it influences the efficiency, but not the site of cDNA integration. Comparisons with other phage DGRs suggested that similar structures are a conserved feature of target sequences. Using a kanamycin resistance determinant as a reporter, we found that translocation of the IMH and hairpin/cruciform-forming region was sufficient to target the DGR diversification machinery to a heterologous gene. In addition to furthering our understanding of DGR retrohoming, our results suggest that DGRs may provide unique tools for directed protein evolution via *in vivo* DNA diversification.

Citation: Guo H, Tse LV, Nieh AW, Czornyj E, Williams S, et al. (2011) Target Site Recognition by a Diversity-Generating Retroelement. *PLoS Genet* 7(12): e1002414. doi:10.1371/journal.pgen.1002414

Editor: William F. Burkholder, Agency for Science, Technology, and Research, Singapore

Received: August 17, 2011; **Accepted:** October 27, 2011; **Published:** December 15, 2011

Copyright: © 2011 Guo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by NIH grants RO1AI071204 and R21DE021528 (JFM). SW was partially supported by grants 1R43AI088979 and R43AI088863 from NIAID. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: JFM is a founder of AvidBiotics Corporation and a member of its Scientific Advisory Board. HG is a consultant of the company. SW is a company employee.

* E-mail: jfmiller@ucla.edu

Introduction

Diversity-generating retroelements (DGRs) have been identified in numerous bacterial phyla [1,2]. Although most DGRs are bacterial chromosomal elements, they are prevalent in phage and plasmid genomes as well. The prototype DGR was identified in a temperate bacteriophage, BPP-1, on the basis of its ability to switch tropism for different receptor molecules on host *Bordetella* species [3]. Tropism switching is mediated by a phage-encoded DGR which introduces nucleotide substitutions in a gene that specifies a host cell-binding protein, Mtd (major tropism determinant), positioned at the distal tips of phage tail fibers. This allows phage adaptation to the dynamic changes in cell surface molecules that occur during the infectious cycle of its bacterial host [3]. Comparative bioinformatics predicts that all DGRs function by a fundamentally similar mechanism using conserved components ([1]; Gingery et al., unpublished data). These include unique reverse transcriptase (RT) genes (*btr* for BPP-1), accessory loci (*axd* or *HRDC*), short DNA repeats, and target genes that are specifically diversified [1–4].

As illustrated by the BPP-1 DGR shown in Figure 1A, diversity results from the introduction of nucleotide substitutions in a

variable repeat (VR) located at the 3' end of the *mtd* gene [1–4]. Variable sites in VR correspond to adenine residues in a homologous template repeat (TR), which remains unchanged throughout the process [1–4]. Transcription of TR provides an essential RNA intermediate that is reverse transcribed by Btr, creating a cDNA product which ultimately replaces the parental VR [4]. During this unidirectional retrotransposition process of mutagenic homing, TR adenines are converted to random nucleotides which subsequently appear at corresponding positions in VR [1–4]. Adenine mutagenesis appears to occur during cDNA synthesis and is likely to be an intrinsic property of the DGR-encoded RT [4].

Located at the 3' end of VR is the IMH (initiation of mutagenic homing) region, which consists of at least two functional elements: a 14 bp GC-only sequence [(GC)₁₄] which is identical to the corresponding segment of TR, and a 21 bp sequence containing 5 mismatches with TR that determines the directionality of information transfer [1]. Using a saturating co-conversion assay, we have precisely mapped a marker transition boundary that appears to represent the point at which 3' cDNA integration occurs and information transfer begins [4]. This maps within the (GC)₁₄ element and we previously postulated that it represents the

Author Summary

Diversity-generating retroelements function through a unique, reverse transcriptase-mediated “copy and replace” mechanism that enables repeated rounds of protein diversification, selection, and optimization. The ability of DGRs to introduce targeted diversity into protein-coding DNA sequences has the potential to dramatically accelerate the evolution of adaptive traits. The utility of these elements in nature is underscored by their widespread distribution throughout the bacterial domain. Here we define DNA sequences and structures that are necessary and sufficient to direct the diversification machinery to specified target sequences. In addition to providing mechanistic insights into conserved features of DGR activity, our results provide a blueprint for the use of DGRs for a broad range of protein engineering applications.

site of a nick or double-strand break in the target DNA [4]. If true, the resulting 3' hydroxyl could serve to prime reverse transcription of the TR-derived RNA intermediate in a target DNA-primed reverse transcription (TPRT) mechanism [4–7]. cDNA integration at the 5' end of VR requires TR/VR homology and may occur via template switching during cDNA synthesis [4].

There are 23 adenines upstream of the (GC)₁₄ element in the BPP-1 TR, each of which is capable of variation [3]. The theoretical maximum DNA sequence diversity is $\sim 10^{14}$, which translates to a maximum protein diversity of nearly 10 trillion distinct polypeptides at the C-terminus of Mtd. For Mtd and other DGR-diversified proteins, co-evolution has resulted in the precise positioning of TR adenines to correspond to solvent exposed residues in the ligand binding pockets of variable proteins [8,9]. As implicated in Figure 1A, mutagenic homing occurs through a “copy and replace” mechanism that precisely regenerates all *cis*-acting components required for further rounds of diversification [4]. This allows the system to operate over and over again to optimize ligand-receptor interactions.

The goal of this study was to characterize requirements for target site recognition by the BPP-1 DGR. Along with insights into the mechanism of mutagenic homing, our results reveal engineering principles that allow DGRs to be exploited to diversify heterologous genes through a process that is entirely contained within bacterial cells.

Results

Boundaries of the BPP-1 DGR target sequence

5' and 3' boundaries of the BPP-1 DGR target sequence were delineated using a PCR-based assay that specifically detects VR sequences that have been modified by DGR-mediated retrohoming [4]. The system consists of a donor plasmid (pMX- Δ TR23-96, Figure 1B) carrying *adv*, a modified TR containing a 30 bp tag (TG2), and *bvt* co-expressed from a BvgAS-regulated promoter [4], and a recipient prophage genome deleted for *adv*, TR, and *bvt* (BPP-1 Δ ATR, Figure 1C). TR retrotransposition from the donor plasmid to the recipient prophage VR creates a “tagged” VR that can be detected using primer pairs specific for the tag and VR-flanking sequences (P1/P4 and P2/P3 in Figure 1B; Table 1). Controls include the demonstration that homing products are Brt-dependent and contain mutagenized adenines. An advantage of this assay is that it does not require infectious phage particle formation and consequently allows manipulation of sequences that are required for Mtd function.

Deletions were introduced into VR and adjacent sequences in BPP-1 Δ ATR lysogens (Figure 1C and Figure S1) and the abilities of mutated prophages to serve as recipients in retrohoming assays were measured (see Materials and Methods). As shown in Figure 1D, sequences upstream of VR were dispensable for DGR homing (lanes 4&13). A deletion mutation that truncates the first 20 bp of VR still supported homing, although at a decreased level (lanes 5&14). Sequence analysis of homing products for this mutant suggested that 5' cDNA integration occurred at cryptic sites within the truncated VR, although 3' cDNA integration occurred in a normal manner (Figure 1D, lanes 5&14; Figure S2). At the 3' end, homing was highly dependent on a 35 bp region located downstream of VR (lanes 6&15 vs. lanes 7&16). This implicated sequences with 8 bp inverted repeats that could potentially form a hairpin structure in ssDNA or a cruciform structure in dsDNA as a possible determinant of DGR target function (Figure 1C). Additional analysis showed that deletion of sequences immediately downstream of the stem was well tolerated (3' Δ 58, Figure 1C and 1E), while further deletions at the 3' end (3' Δ 68) reduced target function to essentially non-detectable levels in homing assays.

In the experiments in Figure 1, homing products were not detected using a donor plasmid expressing enzymatically inactive Brt (BrtSMAA, in which the active site motif YADD is replaced by SMAA; [1,3,4]), and sequence analysis of products generated with primer sets P1/P4 and P2/P3 demonstrated transfer of the TG2 tag from TR to VR. Adenine mutagenesis was observed in $\sim 53\%$ of clones containing P1/P4 products and $\sim 32\%$ of clones containing P2/P3 products (data not shown), which had 3 and 2 TR adenine residues available for mutagenesis, respectively. These observations indicated that true DGR homing products were being detected. Equivalent amounts of template phage DNA, as measured by quantitative PCR, were included in each experiment (lanes 19–27, Figure 1D; lanes 17–24, Figure 1E).

Stem structure, but not sequence, is critical for DGR homing and phage tropism switching

We next determined whether the primary sequence or the secondary structure of the putative hairpin/cruciform located downstream of VR is important for function. To disrupt the structure, 7 consecutive residues proximal to the loop on the 3' half of the stem were changed to their complementary residues (StMut, Figure 2A). The resulting mutant was essentially unable to support DGR homing at a level that could be detected in PCR-based assays (lanes 3&9, Figure 2B). Complementary substitutions were subsequently introduced to the 5' half of the stem to generate StRev (Figure 2A). If the primary sequence is important, the StRev recipient should remain non-functional. Alternatively, if the structure of the stem is the critical element, restoring base pairing interactions might restore DGR target function. As shown in Figure 2B (lanes 5&11), this appears to be the case, as the StRev mutant regained DGR homing activity. Homing products were verified by sequencing and adenine mutagenesis was observed (Figure S3).

Phage tropism switching assays provide a quantitative measure of DGR function [1,3,4]. Although the evolution of new ligand specificities is an inherently stochastic process, the frequency at which it occurs reflects the combined efficiencies of retrohoming and adenine mutagenesis. In Figure 2C, tropism switching was measured using BPP-1 Δ ATR or mutant derivatives complemented with plasmid pMX1, which provides *adv*, TR and *bvt* in *trans* (see Materials and Methods). The StMut mutation resulted in over a 1000-fold decrease in tropism switching, which was restored to near WT levels by the StRev allele. Sequence analysis of VR

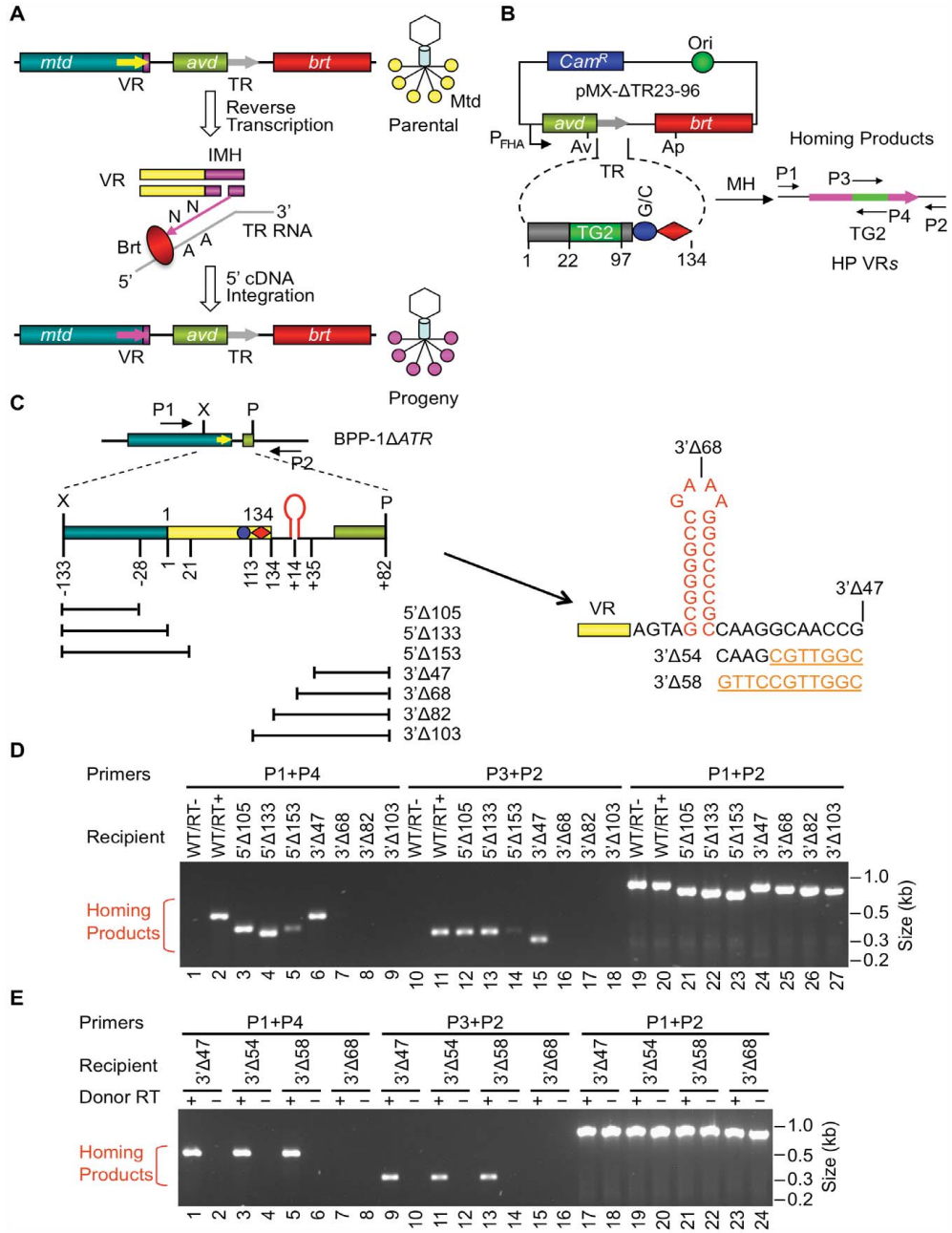


Figure 1. Boundaries of the BPP-1 DGR target sequence. (A) Tropism switching by *Bordetella* phage BPP-1 is mediated by its DGR through mutagenic homing, which is proposed to occur through a target DNA-primed reverse transcription (TPRT) mechanism [4]. *mtd*, *avd*, *brt*, the variable and template repeats (VR and TR) are indicated. VR diversification leads to altered Mtd trimers at the distal tail fibers of progeny phages. (B) PCR-based DGR homing assay. Plasmid pMX- Δ TR23–96 carries the BPP-1 *avd*-TR-*brt* region placed downstream of the BvgA5-regulated P_{Pha} promoter and contains a 30 bp insert (TG2) between TR positions 22 and 97 [4]. Grey and pink arrows represent TR and VR, respectively. Small horizontal arrows indicate primers used for homing assays: P1 and P2 are sense- and antisense-strand primers annealing upstream and downstream of VR, respectively; P3 and P4, sense- and antisense-strand primers, respectively, that anneal to TG2. *Cam*^R, chloramphenicol resistance gene. (C) BPP-1 DGR target region and deletion constructs. 5' or 3' deletions start from position –133 upstream of VR or position +82 downstream of VR, respectively. Lines below the target represent regions deleted. The region downstream of VR contains two 8 bp inverted repeats, which can potentially form a hairpin/cruciform structure. Constructs 3' Δ 54 and 3' Δ 58 were derived from 3' Δ 47 by changing 7 and 11 residues downstream of the hairpin/cruciform structure to their complementary residues, respectively, and were assayed in E. (D) PCR-based DGR homing assays showed that BPP-1 target recognition does not require any sequence upstream of VR, but does require up to 35 bp downstream of VR, which includes the potential hairpin/cruciform-forming region. Donor RT+ indicates pMX- Δ TR23–96, - indicates pMX- Δ TR23–96 with the SMAA mutation. (E) Fine mapping of the 3' boundary of the BPP-1 DGR target with additional deletion constructs 3' Δ 54 and 3' Δ 58 showed that no sequences downstream of position +24 are required for target recognition.
doi:10.1371/journal.pgen.1002414.g001

regions in phages with switched tropisms (5 random clones each) confirmed adenine mutagenesis in every case (Figures S4, S5, S6).

Taken together, these data argue that the ability to form a hairpin or cruciform structure, as opposed to the primary sequence of the inverted repeats, is a critical determinant of target site recognition. The residual tropism switching activity of StMut phage suggests that hairpin/cruciform-independent pathways may exist, although they operate at a much lower efficiency.

Physical evidence for hairpin/cruciform formation in negatively supercoiled DNA

To determine if the hairpin/cruciform structure can form *in vitro*, supercoiled plasmids carrying WT or mutant BPP-1 DGR target sequences were isolated and treated with phage T7 DNA endonuclease I, followed by primer extension with 5' end-labeled primers to identify specific cleavage sites [10,11]. T7 DNA endonuclease I is a structure-specific enzyme that resolves DNA four-way (Holiday) junctions and has previously been used to identify DNA hairpin or cruciform formation [10,11]. As shown in Figure 3, cleavage sites were detected on both DNA strands in the hairpin/cruciform structure, with major cleavage sites at or near the four-way junction. Minor cleavage sites were also detected at or near the loop, as T7 DNA endonuclease I also has some activity on single-stranded DNA [12]. T7 endonuclease I cleavage at the hairpin/cruciform region requires structure formation, as plasmids containing a disrupted stem (StMut) were not cleaved in the corresponding region. Linearization of plasmids containing the WT sequence eliminated cleavage, suggesting that negative supercoiling is required for hairpin/cruciform formation [13,14].

These results demonstrate that hairpins can form on either strand of the target DNA. Although it is likely that they form simultaneously on both strands to create cruciforms, this is not directly addressed by enzyme cleavage assays, hence the hairpin/cruciform designation.

The BPP-1 DGR target sequence functions in an orientation-independent manner

We next determined whether the orientation of the target sequence relative to the phage genome is important for DGR retrohoming. In the experiment in Figure 4A, a segment of the BPP-1 Δ ATR prophage that includes VR and its flanking sequences was inverted, and PCR-based DGR homing assays were performed with donor plasmid pMX- Δ TR23-96. DGR homing into the inverted target occurred at a level comparable to that of the WT control (Figure 4B), and sequence analysis indicated that normal homing products were produced (Figure S7). These results show that the polarity of phage replication is not important for DGR homing, and that the hairpin/cruciform structure functions in a manner that is independent of its orientation relative to the leading or lagging strands formed during DNA replication.

Conservation and functional characteristics of DGR hairpin/cruciform structures

Inverted repeats are nearly always found downstream of VR sequences in target genes [Gingery et al., unpublished data], as illustrated by the phage DGR sequences shown in Figure 5. These elements display a striking pattern of similarity, suggesting they have conserved and important functions. In each case, hairpin/cruciform structures with 7–10 bp GC-rich stems and 4 nt loops can potentially be formed. Although stems are always GC-rich, their sequences differ, while loops are more conserved with the consensus sequence (5'GRNA3', with R = A or G, N = any nucleotide) in the sense strand. The exact distance between the hairpin/cruciform structures and the 3' ends of their respective VRs appears to be quite flexible. We took advantage of the BPP-1 DGR system to test the relevance of these patterns of conservation, with the goal of generating a more comprehensive understanding of parameters important for target site recognition.

We first studied requirements for stem length and sequence and found that although minor changes are tolerated, the WT configuration appears to be optimized for BPP-1 DGR function. Of the stem length variants in Figure 6A, extensions are better tolerated than deletions. Removal of 2, 4 or 6 bp proximal to the loop results in markedly decreased activity in both PCR-based homing (Figure 6B) and phage tropism switching assays (Figure 6C), to levels similar to those observed with the StMut allele in which the stem is completely abolished (Figure 2A). Insertion of 2 bp next to

Table 1. List of oligonucleotides used in this study.

Name	Sequence
P1	5' TTCGGTACCTGCTAGCGTCAACACCTG
P2	5' AGCAAGCTTGCTCTGTTGCGGTGATGCT
P3	5' AAATCTAGATCTGCTGCGTTTGTT
P4	5' AGCAAGCTTAGCACAGAACACAAACG
P5	5' GGTACCATGAGCATTGGTCGTAGCA
P6	5' GTACAGCGGGCCGTCGTTCTCGTTCGCGTT
P7	5' CCCTCTAGAGCTCCGGTTGCTTGTGGACG
P8	5' AGCAAGCTTCTCGATGGGTTCCAT
P9	5' ATATCTAGACGTTTTCTGGGCTACCCGTTTAATGTCG
P10	5' ATAAAGCTTCGACATTAACCGGTAGACCCCAAGAAAA

doi:10.1371/journal.pgen.1002414.t001

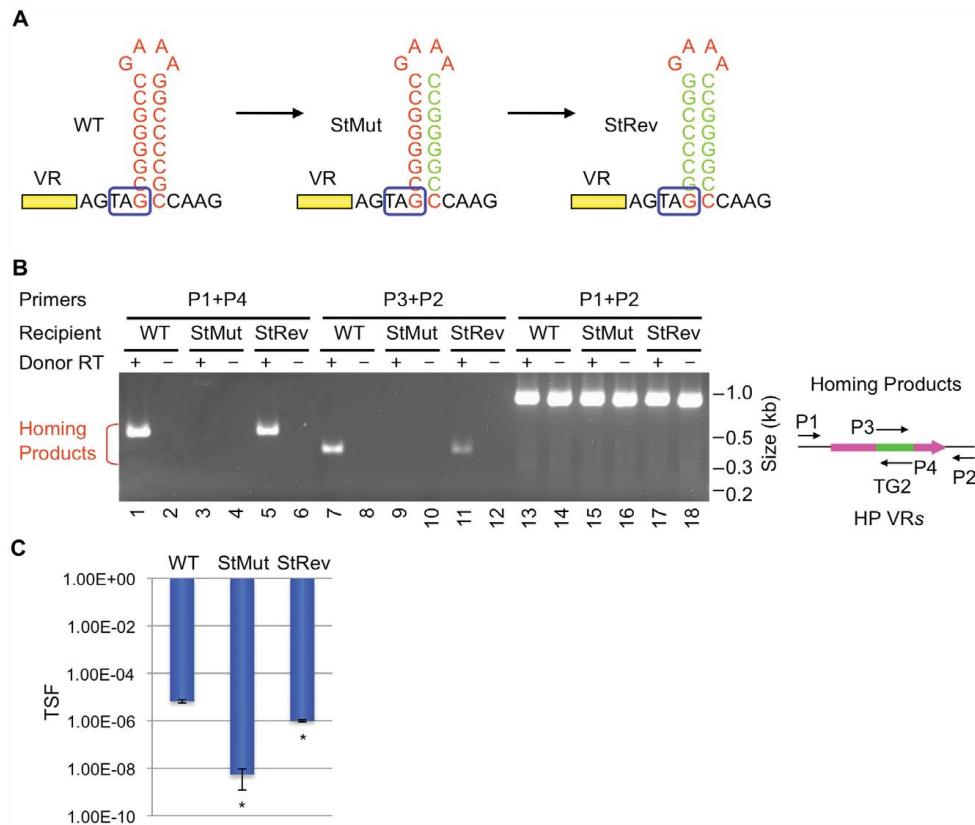


Figure 2. The hairpin/cruciform structure downstream of VR is required for target recognition. (A) WT and mutant hairpin/cruciform structures. Boxed TAG, *mtd* stop codon. (B) The hairpin/cruciform structure downstream of VR, as opposed to the primary sequence of the stem, is required for DGR mutagenic homing. PCR-based DGR homing assays are shown on the left, with primers used to detect homing products shown on the right. Sequence analysis of homing products demonstrated TG2 transfer from TR to VR with WT (data not shown) and StRev recipients (Figure S3). Adenine mutagenesis was observed in 7/12 of WT and 8/15 of StRev homing products amplified with primers P1 and P4, and 6/10 of WT and 5/20 of StRev homing products amplified with P2 and P3 (data not shown and Figure S3). (C) The hairpin/cruciform structure downstream of VR is required for phage tropism switching. The graph shows phage tropism switching frequencies (TSF) for BPP-1 Δ ATR with WT or mutant hairpin/cruciform structures. The bars represent mean TSF \pm standard deviations (s.d.). P values comparing mutants to WT in Student's t tests are indicated with asterisks. * $P < 0.02$. doi:10.1371/journal.pgen.1002414.g002

the loop had little effect on activity, while longer insertions gradually decreased target site function. Keeping the length of the stem constant, a sequence change in the middle of the stem that converts 4 GC base pairs to AT base pairs (StAT, Figure 6A) greatly reduced, but did not eliminate function. We next tested the effects of altering the sequence and size of the loop using the mutant constructs shown in Figure 7A. Substituting CTTT for the consensus loop sequence GAAA, or increasing the size of the loop by as little as 2 nt, decreased activity in PCR-homing (Figure 7B) and tropism switching assays (Figure 7C) to near background levels. Based on these experiments, it appears that an 8–10 bp GC-rich stem is optimal for BPP-1 DGR homing, and that both the size and sequence of the 4 bp loop are critical for function. Our results correlate with the patterns of conservation shown in Figure 5.

Shifting the position of the hairpin/cruciform element alters the efficiency of target site recognition but not the site of 3' cDNA integration

In the experiments in Figure 8, we tested the effects of altering the position of the hairpin/cruciform with respect to the 3' boundary of VR and probed sequence requirements for the intervening region. SpM4 (Figure 8A), in which the 4 residues in the spacer were switched to the complementary nucleotides, retained WT activity (Figure 8B). In contrast, deletion of the spacer (SpD4) resulted in a significant decrease in target function. The SpM4 and SpD4 mutations eliminate the *mtd* stop codon and generate non-infective phages, obviating the ability to measure tropism switching. Nonetheless, their relative levels of activity were readily apparent in PCR-homing assays. Expansion of the spacer

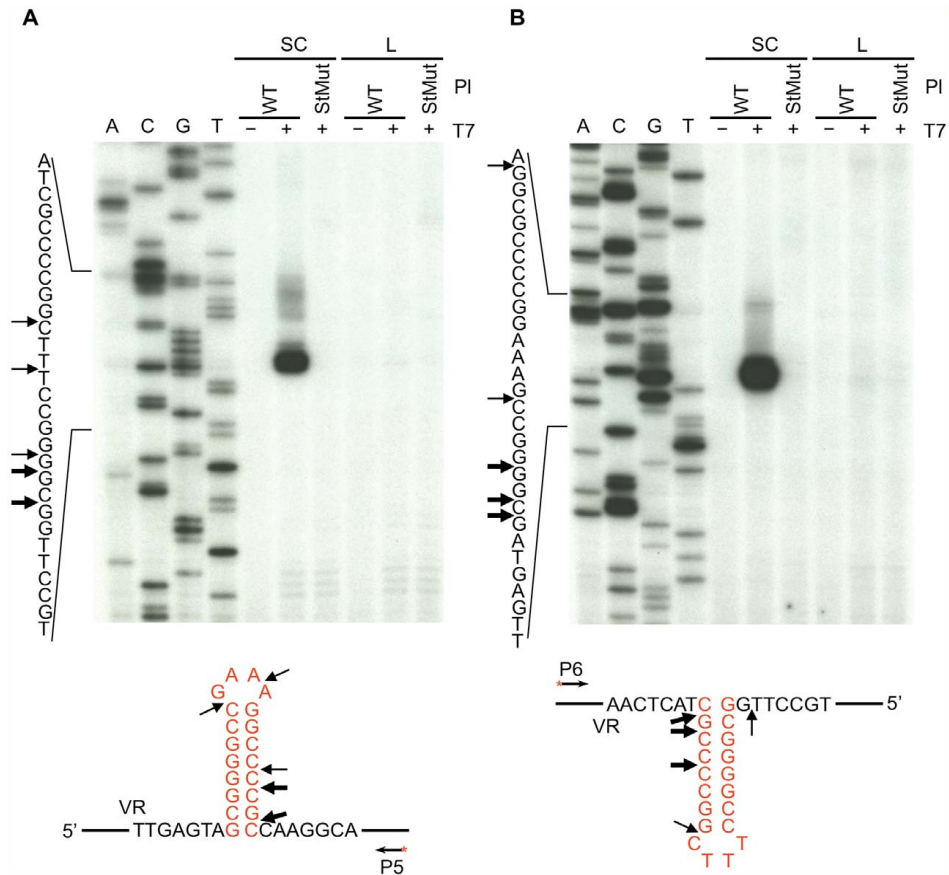


Figure 3. Hairpin/cruciform structure formation in negatively supercoiled DNA. Primer extension assays were used to identify T7 endonuclease I cleavage sites in plasmid DNA containing either the WT or StMut target sequences. (A) Top strand cleavage, (B) bottom strand cleavage. Supercoiled or linearized plasmid DNA was either left untreated (–) or digested with T7 endonuclease I (+) followed by primer extension. Sequence ladders are shown to the left and primer extension termination sites in supercoiled WT plasmid DNA are shown underneath the gels. Thick and thin arrows designate major and minor cleavage sites, respectively. The exact positions of cleavage sites are ± 2 –3 nt due to uncertainty resulting from compression of the sequence ladder in the hairpin/cruciform region caused by DNA secondary structure. The heavily labeled major products on both strands represent multiple adjacent cleavage sites which were resolved at lower exposures. The figure is representative of results obtained from multiple independent experiments. P5 and P6 (Table 1) are the 5' end-labeled primers used for primer extensions in A and B, respectively. SC, supercoiled plasmids; L, linearized plasmids; PI, plasmid; T7, T7 endonuclease I. doi:10.1371/journal.pgen.1002414.g003

was tolerated to a greater extent than deletion. SpI3, which has a 3 bp insertion in the spacer (Figure 8A), showed no significant defect in PCR-homing or phage tropism switching assays (Figure 8B and 8C), but longer insertions gradually decreased target site function.

The SpI6 insertion, which increases the distance between the hairpin/cruciform structure and the 3' end of VR by 6 bp, retains a measurable level of activity. We took advantage of this and used a marker coconversion assay (Figure S8; [4]) to determine the relationship between the position of the hairpin/cruciform structure and the site at which information transfer initiates. As

summarized in Figure 8D, our coconversion assay measured transfer of nucleotide polymorphisms from tagged TR donors to a recipient VR carrying the SpI6 mutation using PCR-based homing assays (data not shown). With the WT recipient, a coconversion boundary occurs between positions 107 and 112, and this was interpreted as representing the site at which TR-derived cDNA synthesis initiates [4]. As shown in Figure 8D, the coconversion boundary remains essentially unchanged in the SpI6 mutant. Although the position of the hairpin structure affects the efficiency of DGR homing, it does not determine the site at which cDNA is integrated at the 3' end of VR.

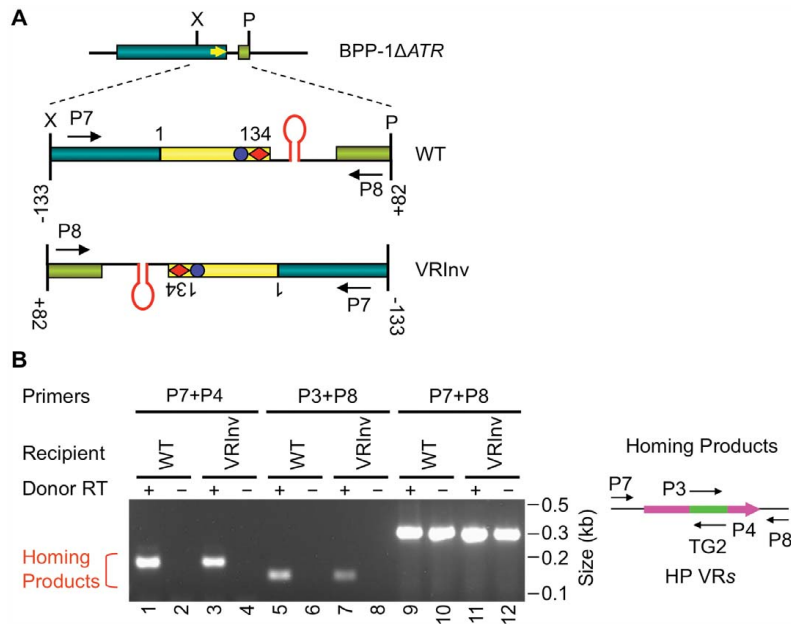


Figure 4. Target orientation in the BPP-1 phage genome is not critical for recognition. (A) Schematic of the BPP-1ΔATR phage genome with WT or inverted (VRInv) target sequences. A segment from position -133 upstream of VR to position +82 downstream of VR was inverted. VR-flanking primers used in the DGR homing assay are also indicated. (B) Target inversion had no significant effect on DGR homing activity. pMX-ΔTR23-96 or its RT-deficient derivative were used as donors for the indicated recipients and products from PCR-homing assays are shown. The diagram to the right shows primers used for the assay. Sequence analysis of VRInv homing products showed TG2 transfer from TR to VR, as well as adenine mutagenesis in 5/10 homing products amplified with primers P7 (Table 1) and P4, and 2/9 products amplified with P8 (Table 1) and P3 (Figure S7). doi:10.1371/journal.pgen.1002414.g004

Engineering the BPP-1 DGR to target a heterologous gene

To determine if the results presented here complete our understanding of DGR-encoded requirements for retrohoming to a target gene, we applied them as engineering principles in an attempt to construct a functional, synthetic, TR/VR system. For a DNA sequence to serve as a recipient VR, three conditions must be met. First, it must be adjacent to an IMH region with functional (GC)₁₄ and 21 bp elements at its 3' end [1,4]. Second, the IMH region must be followed by inverted repeats capable of forming a hairpin/cruciform structure of appropriate size, composition and distance from IMH. And finally, sufficient VR/TR sequence homology must be provided to allow efficient upstream (5') cDNA integration. In recent studies we have shown that although short stretches of nucleotide identity (≥ 8 bp) between the TR-derived cDNA and VR target sequences are sufficient to complete the homing reaction, homing efficiency is increased with longer (≥ 19 bp) stretches of homology [4]. With these parameters in mind, we tested our ability to engineer the BPP-1 DGR to target a heterologous reporter gene (*aph3' Ia*; [15]) which provides facile detection of targeting events by antibiotic selection.

The recipient VR-*Kan^S* cassette shown in Figure 9A contains an *aph3' Ia* kanamycin resistance (*Kan^R*) allele with a 3' deletion that renders it nonfunctional by removing coding sequences for 6 essential C-terminal residues. The truncated gene was placed immediately upstream of IMH, followed by the hairpin/

cruciform-forming inverted repeats from the BPP-1 DGR. Transcription is directed by the native *aph3' Ia* promoter. The donor plasmid expresses *avd*, *bvt*, and one of two engineered TRs (TR-Km1, TR-Km2) from the *P_{ma}* promoter. Both TRs contain the intact 3' end of the *aph3' Ia* open reading frame, followed by two consecutive stop codons and sequences 97–134 from the 3' end of the BPP-1 TR. For TR-Km2, the *aph3' Ia* fragment is also flanked, at its 5' end, by the first 22 residues of the BPP-1 TR. DGR-mediated retrotransposition from the donor TR constructs to the VR-*Kan^S* recipient should regenerate a full-length *aph3' Ia* gene conferring *Kan^R*.

We first tested whether targeting can occur in the context of a replicating phage. BPP-1ΔATR^{Kan^S} carries the VR-*Kan^S* cassette inserted between *attL* and *bhb1* on the left arm of the prophage genome [16], along with a deletion of *avd*, TR and *bvt* and a series of synonymous substitutions in IMH to inactivate the *mtl* VR (Figure 7A). *B. bronchiseptica* RB50 carrying the TR-Km1 or -Km2 donor plasmid, or derivatives with a null mutation in *bvt*, were infected with BPP-1ΔATR^{Kan^S} and targeting efficiencies were determined by infecting RB50 with progeny phages and measuring relative numbers of Kan^R lysogens. Kan^R lysogens were readily detected when targeting occurred from Brt+ TR donors, but not Brt- donors (Figure 9B), and sequence analysis showed that Kan^R resulted from the regeneration of full-length *aph3' Ia* alleles which often contained mutations at positions corresponding to adenines in donor TRs (Figures S9 and S10).

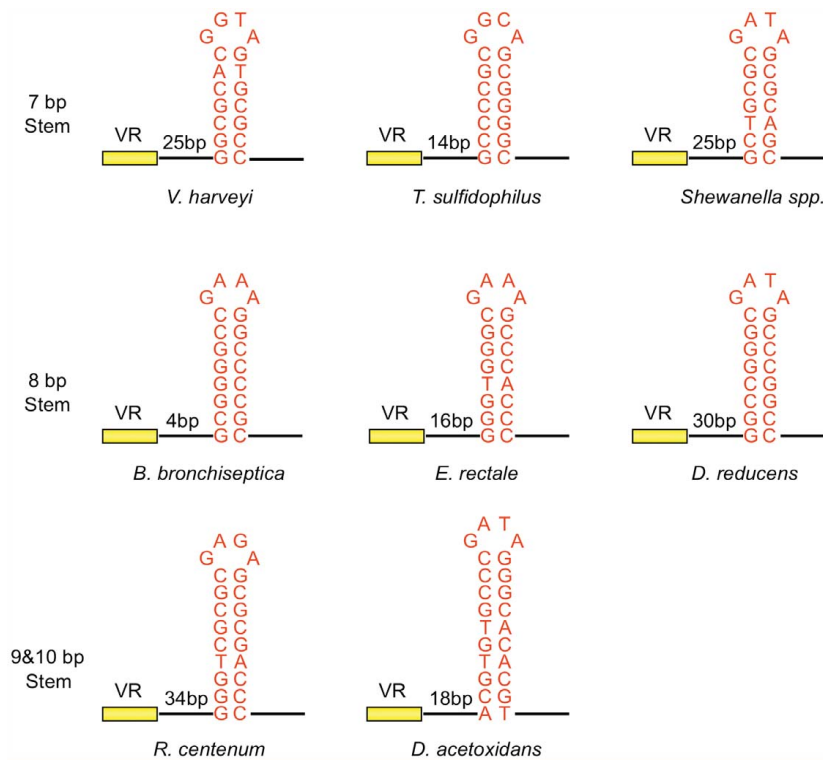


Figure 5. Phage-related DGRs contain potential hairpin/cruciform structures with conserved features. DGRs associated with phages or phage-related sequences in different bacterial genomes were identified as described in Doulatov *et al.* [1]. Short inverted repeats which could potentially form hairpin/cruciform structures were found downstream of VRs as shown. In each case, GC-rich stems are 7–10 bp in length with 4 nt loops composed of the conserved sequence [5'GRNA; R= A or G, N= any nucleotide]. Relative distances between the hairpin/cruciform structures and their corresponding VRs range from 4 to 34 bp. *V. harveyi*, *Vibrio harveyi* phage VHML; *T. sulfidophilus*, *Thioalkalivibrio sulfidophilus* HL-EbGr7; *Shewanella sp.*, *Shewanella sp.* W3-18-1; *B. bronchiseptica*, *Bordetella bronchiseptica* phage BPP-1; *E. rectale*, *Eubacterium rectale* DSM 17629; *D. reducens*, *Desulfotomaculum reducens* M1-1; *R. centenum*, *Rhodospirillum centenum* SW; *D. acetoxidans*, *Desulfotomaculum acetoxidans* DSM 771. doi:10.1371/journal.pgen.1002414.g005

It is interesting to note that the TR-Km1 donor was significantly more efficient than TR-Km2. This suggests that the majority of cDNAs are extended to the 5' termini of these short synthetic TRs, and target (VR) homology to the extreme 3' ends of the extension products may be advantageous for cDNA integration.

We also tested the ability to target the VR-*Kan^S* cassette when present on a resident prophage in the bacterial chromosome or on a plasmid. In the experiment in Figure 9C, RB50/BPP-1Δ*ATR^RKan^S* lysogens were transformed with donor plasmids under conditions that suppress P_{pha} promoter activity. Following a 6 hr pulse of P_{pha} induction, cells were plated under promoter-suppressing conditions on media with or without kanamycin. In Figure 9D, a similar protocol was used to target a VR-*Kan^S* cassette carried on a medium copy number plasmid in RB50 cells containing a TR donor plasmid, but no other phage sequences. In both experiments, Kan^{R} colonies were readily detected when targeting occurred from Brt^{+} , but not Brt^{-} TR donors, and sequence analysis showed characteristic patterns of adenine mutagenesis (Figures S11, S12, S13, S14). Taken together, our

results demonstrate the ability to engineer a VR/TR system that targets a heterologous reporter gene on a phage, plasmid or bacterial genome. The data in Figure 9D show that no BPP-1 phage products, other than those encoded in the DGR, are required for mutagenic retrohoming.

Discussion

Understanding DGR target site recognition requires a precise definition of *cis*-acting sequences important for retrohoming. Our analysis of the boundaries of the BPP-1 DGR target showed that sequences upstream of VR are dispensable, as predicted by previous results [4]. More importantly, we show that homing is facilitated by an element downstream of VR, beyond the point at which TR/VR homology ends. Sequence analysis, mutagenesis, and structure-specific nuclease assays demonstrated that GC-rich inverted repeats directly following VR form a hairpin/cruciform structure that plays a critical role in retrohoming. Highly similar elements are present in analogous locations in many phage- or

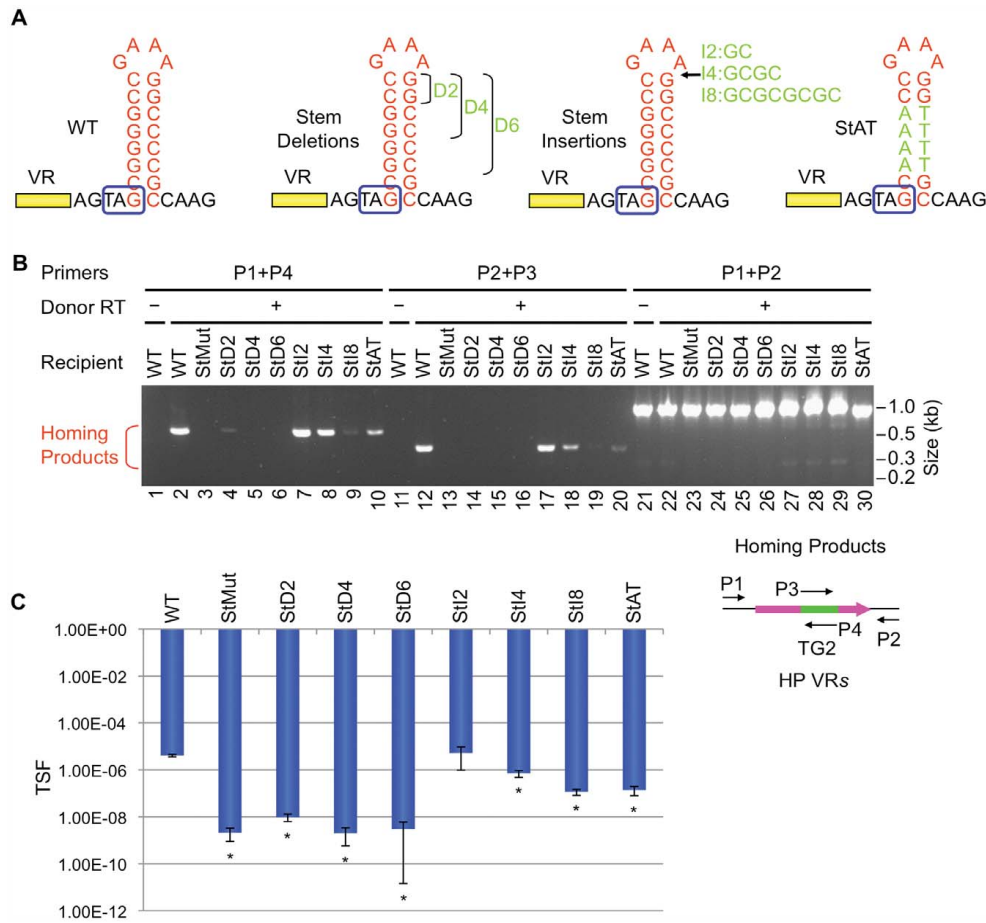


Figure 6. The stem length and sequence are important but can tolerate minor modifications. (A) Modifications in the stem of the hairpin/cruciform structure are shown. StD2, StD4 and StD6 are deletion mutants; StI2, StI4 and StI8 are insertion mutants. StAT has four GC base pairs in the middle of the stem changed to AT base pairs. Boxed TAG, *mtd* stop codon. (B) The stem is important for BPP-1 DGR mutagenic homing but can tolerate minor changes. PCR-homing assays are shown along with primers used. (C) Effects of stem modifications on phage tropism switching frequencies. The scale bars represent mean TSF \pm s.d. P values comparing mutants to WT in Student's t tests are indicated with asterisks. * $P < 0.02$. doi:10.1371/journal.pgen.1002414.g006

prophage-related DGRs (Figure 5), and hairpin/cruciform structures are predicted for the majority of DGRs that naturally reside on bacterial chromosomes and plasmids as well [Gingery et al., unpublished data]. We propose that DNA hairpin formation near the 3' end of VR is a conserved requirement for DGR-mediated retrohoming.

For the BPP-1 DGR target, the 8 bp stem appears to function as a structure that is dependent on nucleotide composition but not sequence. In contrast, the loop of the hairpin/cruciform structure is constrained in size and sequence and conforms to the consensus, 5'-GRNA, derived from comparisons with other phage-related DGRs. This suggests that loop sequence and size may be important for stabilizing the hairpin/cruciform structure [17], or

for creating a strand bias in DNA cleavage by a host-encoded endonuclease. It is also possible that the loop is in direct physical contact with a critical component, such as Brt, Avd, a TR-containing RNA transcript, or other parts of the DGR target. By testing the effects of length and sequence variations between the hairpin/cruciform and VR, we found that distance is an important parameter, although some flexibility exists. Extending the spacer by 6 bp did not shift the marker coconversion boundary in the (GC)₁₄ region during DGR homing [4], showing that the position of the hairpin/cruciform does not determine the site at which 3' cDNA integration occurs.

DGRs are evolutionarily related to group II introns [1] and it is interesting to note that a subset of these retroelements, the

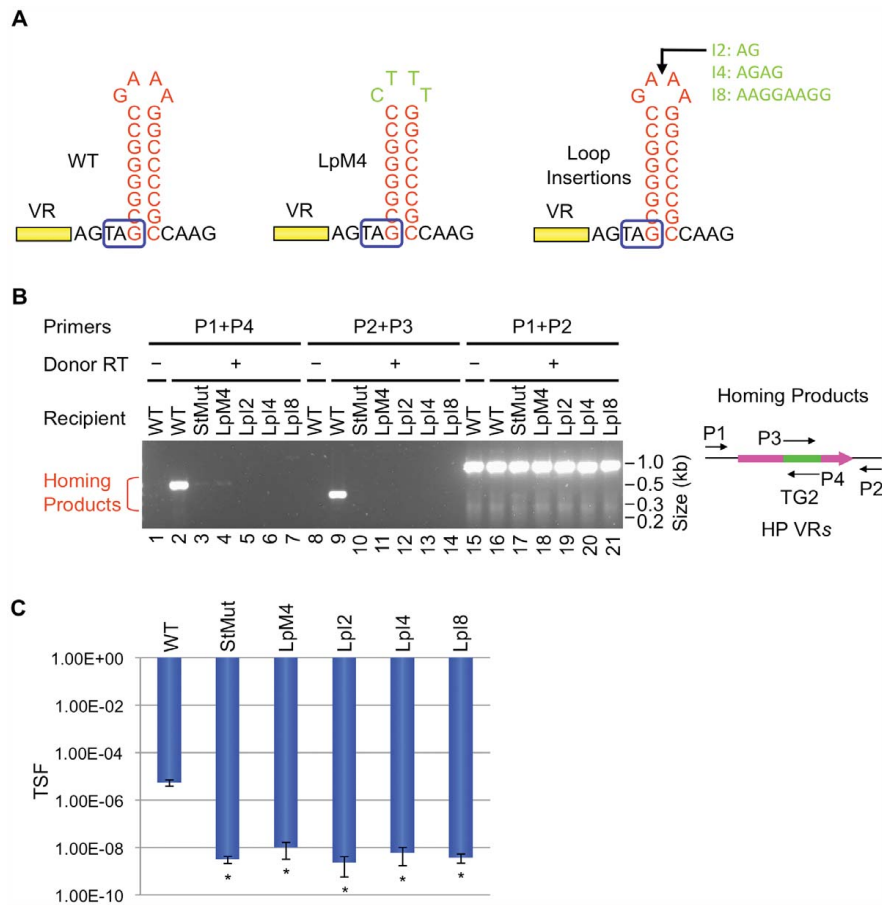


Figure 7. The loop of the hairpin/cruciform structure is critical for DGR function. (A) Substitution mutations or insertions in loop sequences are shown. Boxed TAG, *mtd* stop codon. (B) Both the loop sequence and size are highly critical for DGR mutagenic homing. PCR-homing assays are shown along with primers used. (C) Effects of loop modifications on phage tropism switching frequencies. The bars represent mean TSF \pm s.d. P values comparing mutants to WT in Student's t tests are indicated with asterisks. * $P < 0.05$. doi:10.1371/journal.pgen.1002414.g007

group IIC introns, also target motifs with stem-loop structures [18–20]. In nature, group IIC introns are often found to be located short distances downstream of sequences encoding known or predicted factor-independent transcription terminators, which are composed of GC-rich stems with loops of varying sizes followed by poly-uridine stretches [18–20]. Using an *in vitro* mobility assay, Robart *et al.* [19] have shown that reconstituted ribonucleoprotein particles from the *Bacillus halodurans* *B.h.II* group IIC intron recognize structures in ssDNA that correspond to RNA hairpins formed during transcription termination. As observed with the BPP-1 DGR, the *B.h.II* mobility reaction was highly dependent on stem formation but not absolute sequence [19]. Stems shorter than

9 bp had significantly reduced activities in *in vitro* mobility assays, a longer stem (14 bp) retained function, and the efficiency of targeting correlated with GC content and predicted stem stability [19]. In contrast to our observations with the BPP-1 DGR, alterations in loop sequence had little effect on *B.h.II* mobility *in vitro* [19]. The adaptation of group IIC introns to recognize and insert downstream of factor-independent transcriptional terminators was proposed to provide a selective advantage by limiting their expression, avoiding the interruption of essential coding sequences, and facilitating horizontal spread as intrinsic terminators are common and conserved in bacteria [19]. For DGRs, we speculate that the ability to target sequences upstream of

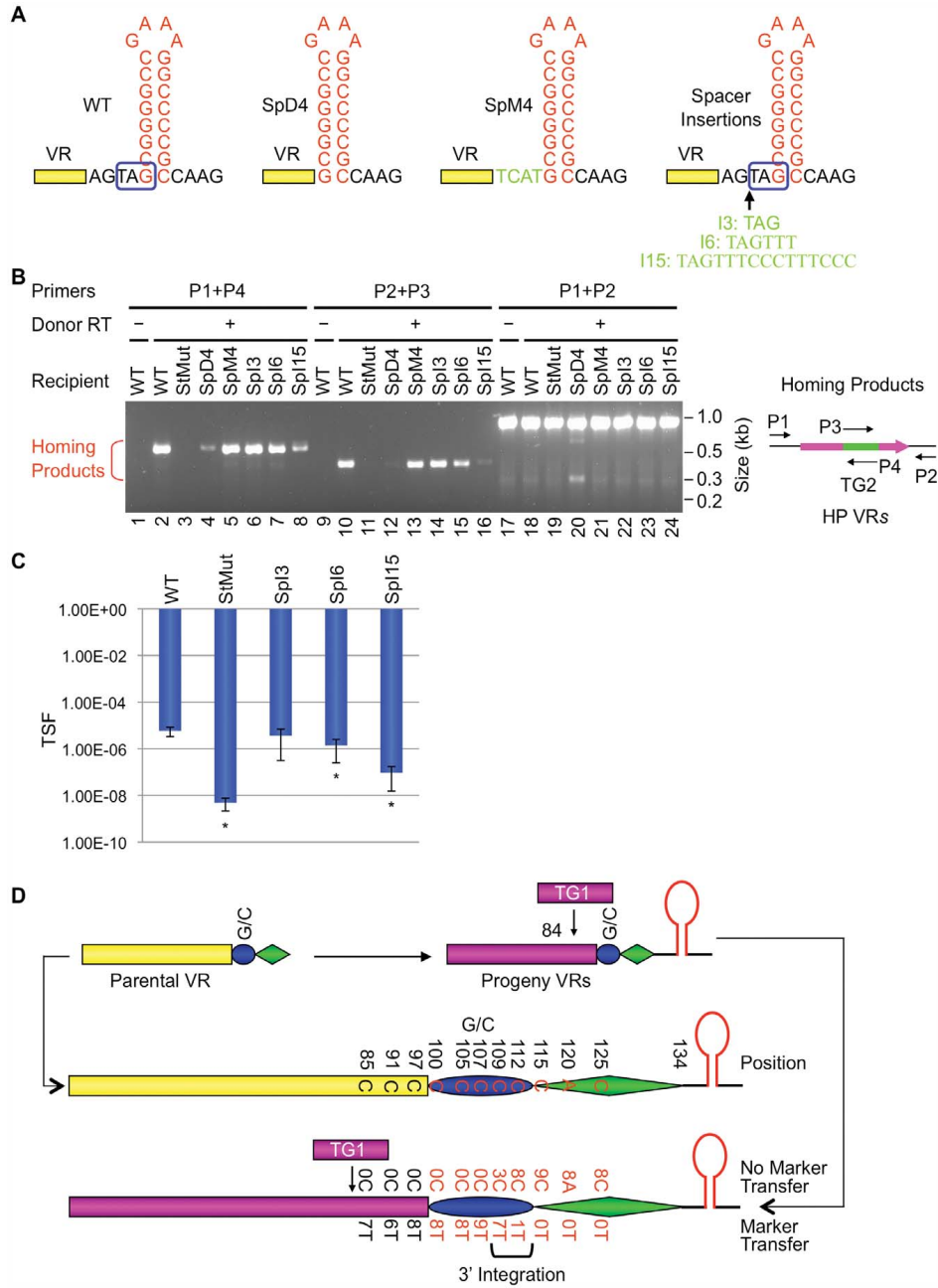


Figure 8. The VR-hairpin/cruciform spacer affects target recognition efficiency but not 3' cDNA integration site. (A) Modifications introduced in the spacer between VR and the hairpin/cruciform structure. SpD4, the 4 bp spacer was deleted; SpM4, residues of the 4 bp spacer were switched to their complementary nucleotides; SpI3, SpI6 and SpI15 are spacer insertion mutants that retain the *mtd* stop codon for proper Mtd production. Boxed TAG, *mtd* stop codon. (B) Effects of spacer modifications on DGR mutagenic homing. PCR-homing assays are shown along with primers used. (C) Effects of spacer modifications on phage tropism switching. The bars represent mean TSF \pm s.d. P values comparing mutants to WT in Student's t test are indicated with asterisks. *P<0.05. Tropism switching data are not available for constructs SpD4 and SpM4, as both modifications eliminate the *mtd* stop codon and do not produce functional phage particles. (D) Summary of marker coconversion analysis with BPP-1 Δ ATRSpI6. BPP-1 Δ ATRSpI6 phage particles were used for single-cycle lytic infection of RB50 cells transformed with marked donor plasmids. Progeny phage DNAs were used for PCR-based DGR homing assays. PCR products were cloned and sequenced and marker coconversion data are summarized. Nucleotide residues in the parental VR that correspond to the marked positions in TR are shown in the center. Numbers of progeny VRs with or without transferred markers at designated sites are shown at the bottom. The deduced cDNA integration region at the 3' end of VR is indicated by a bracket.

doi:10.1371/journal.pgen.1002414.g008

terminator-like stem-loop structures may have played a role in directing their sequence diversification capabilities to the 3' coding regions of target genes.

The TPRT model for DGR homing postulates that cDNA synthesis initiates with a nick or double-strand break in the IMH (GC)₁₄ sequence, providing a primer for reverse transcription of a TR-containing RNA transcript [4]. Analogous to target recognition by group IIC introns, the hairpin/cruciform structure may serve as a recognition element for a retrohoming complex that includes *trans*-acting DGR-encoded factors. A DNA endonuclease that might be responsible for cleavage awaits identification, and possibilities include Avd, Brt, a TR-derived catalytic RNA, or an unidentified host factor. It is also possible that the DNA hairpin/cruciform actively promotes single- or double-strand breaks. If DNA repair synthesis extends to the (GC)₁₄ region, the elongating antisense strand could then be used for cDNA priming. DNA breaks at the hairpin/cruciform structure could be created by an endonuclease that cleaves the single-stranded loop, or by a structure-specific enzyme similar to T7 endonuclease I [21]. Since DNA cruciforms are structurally similar to Holiday junctions, host-encoded recombination proteins that function in resolving recombination intermediates could be involved [22]. The cDNA priming mechanism of the BPP-1 DGR appears to be different from that of mobile group II introns that lack a DNA endonuclease activity in their intron-encoded proteins [23–25]. Reverse transcription in retrohoming and ectopic transposition of these elements is proposed to be primed by either the leading or lagging strand during DNA replication, and strong strand-specific biases are observed [23–25]. Our observation that the BPP-1 DGR target sequence is orientation-independent suggests that DNA replication polarity does not play a significant role in cDNA priming. Although our results to date are consistent with TPRT, further studies are required to definitively characterize the mechanism of cDNA initiation and integration at the 3' end of VR and to determine the precise role of the hairpin/cruciform structure in the retrohoming process.

The broad distribution of DGRs in nature attests to their utility, and prospects for adapting these elements for protein engineering applications are compelling. Our results demonstrate that the region containing the (GC)₁₄ and 21 bp sequences in IMH, and an adjacent hairpin/cruciform, is sufficient to direct the DGR mutagenic homing machinery to a heterologous target gene through appropriate engineering of a cognate TR. Using similar design principles we have successfully targeted a tetracycline resistance determinant as well (HG and JFM, unpublished data). For DGRs to be useful tools, it will be necessary to engineer their activity to allow efficient and controlled diversification. Having defined the DGR-encoded *cis*- and *trans*-acting factors required to diversify heterologous sequences, efforts to optimize their activities can now proceed in an informed and comprehensive way. It will also be important to determine the effects of TR/VR size,

composition, and position relative to *cis*-acting DGR elements, on the efficiency of diversifying heterologous sequences. In preliminary experiments, insertions of moderate size (up to ~200 bp) at position 84 in the BPP-1 TR (134 bp) are transferred to VR and mutagenized at adenines, suggesting that sequences of >300 bp could be diversified by an engineered system (LVT, HG and JFM, unpublished data).

In addition to providing prodigious levels of diversity, mutagenic homing is a regenerative process that allows DGRs to operate through unlimited rounds to optimize variable protein functions [4]. 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, through a process that takes place entirely within bacterial cells.

Materials and Methods

Bacterial strains and phages

B. bronchiseptica strains RB50, RB53Cm, RB54 and ML6401 have been described [16]. The BPP-1 Δ ATR lysogen was constructed from ML6401, an RB50 strain lysogenized with phage BPP-1, by deleting sequences from *adv* position 48 to position 882 of *brt*. Target region deletions/insertions and hairpin/cruciform modifications were introduced into the BPP-1 Δ ATR lysogen through allelic exchange [1,4] and are diagrammed in the figures. The BPP-1 Δ ATR* lysogen contains multiple silent mutations at both the 5' and 3' ends of VR to inactivate it as a DGR target. It was used as the parental strain to create the BPP-1 Δ ATR**Kan*^R lysogen, in which the *Kan*^R gene *aph3' Ia* has sequences encoding the C-terminal 6 amino acid residues truncated and is placed upstream of IMH and the hairpin/cruciform structure as a reporter for heterologous gene targeting. The *aph3' Ia* allele also contains an AAA to CGC substitution resulting in K260R. The VR-*Kan*^S reporter cassette was inserted between *attL* and *bbp1* of the phage genome. Phage BPP-1 Δ ATR and its various derivatives were produced from the above lysogens.

Plasmid constructs

Plasmid pMX- Δ TR23–96 has TR positions 23–96 deleted and replaced by a 30 bp PCR tag as in pMX- Δ TR23–84 [4]. Its RT-deficient derivative contains the YMDD to SMAA mutation at Brt positions 213–216 [3,4]. Plasmids pMX1 and pMX1SMAA were used for phage tropism switching assays and have previously been described [4].

pUC-StWT is a pUC18-based plasmid containing the WT BPP-1 DGR target from position –6 upstream of VR to position +82 downstream of VR. pUC-StMut is its derivative with 7 residues in the 3' half of the stem, proximal to the loop, mutated to their complementary nucleotides.

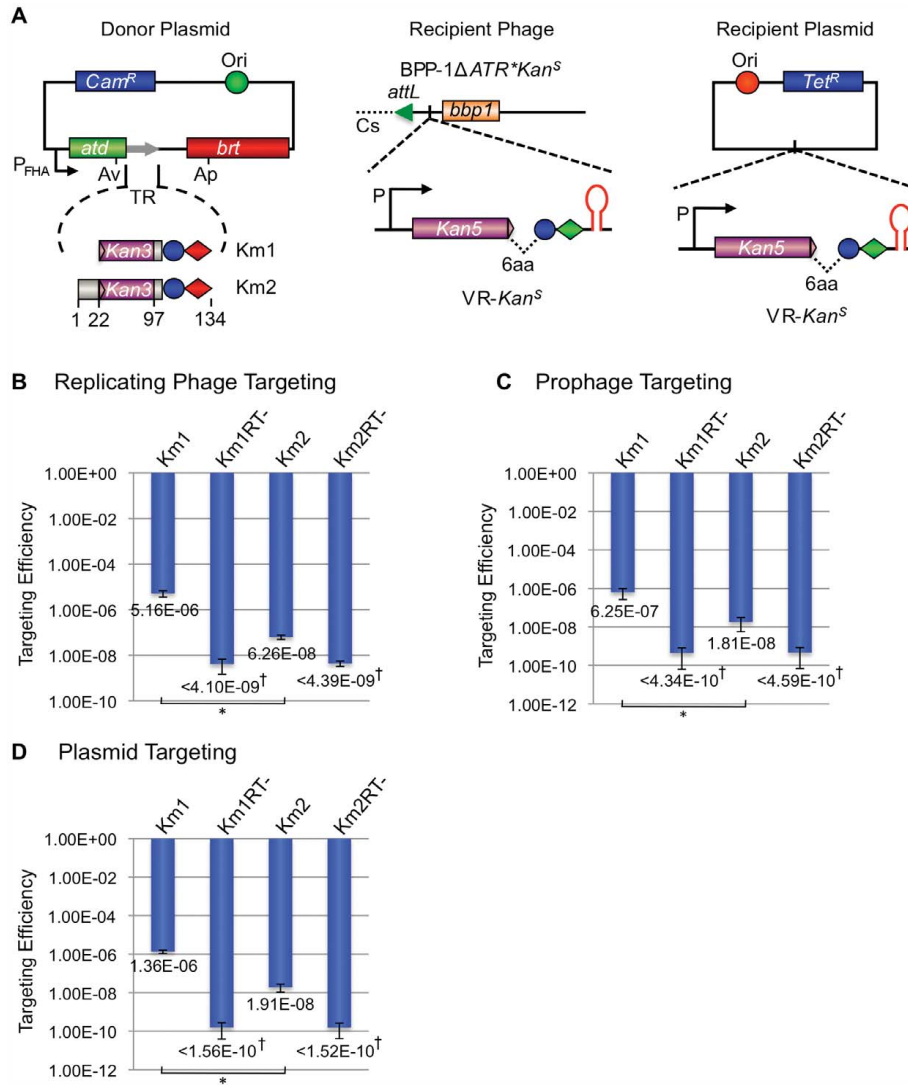


Figure 9. The BPP-1 DGR can be engineered to target a kanamycin-resistance gene. (A) Donor and recipient constructs used for *Kan^R* gene targeting. The left panel shows two donor plasmids, pMX-Km1 and pMX-Km2, both with engineered TRs containing the last 36 bp of the *Kan^R* ORF. The *Kan^R* reporter gene on recipient VRs has a truncation of the last 6 codons and is placed upstream of the IMH and hairpin/cruciform elements. The recipient cassette was inserted between *attL* and *bbp1* of phage BPP-1Δ*ATR** (Recipient Phage, middle) or in a pMMB208-derived plasmid (Recipient Plasmid, left). (B) Engineered BPP-1 DGRs can target a *Kan^R* reporter gene on a replicating phage. Targeting assays were carried out by phage BPP-1Δ*ATR***Kan^S* lytic infection of RB50 cells carrying the indicated donor plasmids. Targeting efficiencies were determined as the relative numbers of *Kan^R* cells in lysogens generated with fresh RB50 cells and the progeny phages. The bars for Km1 and Km2 represent mean ± s.d. **P*<0.01 in a Student's t test. (C) Engineered BPP-1 DGRs can target a *Kan^R* reporter gene on a prophage in the bacterial chromosome. BPP-1Δ*ATR***Kan^S* lysogens transformed with indicated donor plasmids were used for targeting assays. Resulting cells were directly plated on selective (+Kan) or non-selective plates to determine relative numbers of *Kan^R* cells. The bars represent mean ± s.d. **P*<0.05 in a Student's t test. (D) Engineered BPP-1 DGRs can target a *Kan^R* reporter gene on a plasmid. Following induction for donor plasmid expression and *Kan^R* gene targeting, RB50 cells transformed with both donor and recipient plasmids were plated on selective (+Kan) or non-selective plates to determine targeting efficiencies as in (C). The bars represent mean ± s.d. **P*<0.002 in a Student's t test. [†]No *Kan^R* colonies were observed for RT-deficient donors and the numbers represent limits of detection. doi:10.1371/journal.pgen.1002414.g009

Plasmids pMX-TRC85T, pMX-TRC91T, pMX-TRC97T, pMX-TRC100T, pMX-TRC105T, pMX-TRC107T, pMX-TRC109T, pMX-TRC112T, pMX-TRC115T, pMX-TRC120T and pMX-TRC125T have been previously described [4].

Plasmids pMX-Km1 and pMX-Km2 were constructed from pMX- Δ TR23–96 for *Kan^R* gene targeting, both containing the last 36 bp of *aph3' Ia*. The 36 bp sequence and its following two stop codons replace TR positions 1–96 in pMX-Km1 and TR positions 23–96 in pMX-Km2.

Plasmid pHGT-*Kan^S* contains the VR-*Kan^S* cassette described above and was used as the recipient plasmid for *Kan^R* targeting. The plasmid also carries a tetracycline resistance gene.

Phage production for DGR homing and tropism switching assays

Phage production for DGR functional assays was carried out by either single-cycle lytic infection or mitomycin C induction from lysogens as previously described [4], except for minor modifications as noted. For single-cycle lytic infection, *B. bronchiseptica* RB50 cells transformed with appropriate donor plasmids were grown overnight at 37°C in Luria-Bertani (LB) media containing 25 µg/ml of chloramphenicol (Cam), 20 µg/ml streptomycin (Str), and 10 mM nicotinic acid to modulate to the Bvg⁺ phase and prevent transcription from the P_{tha} promoter. An amount of cells equal to 1 ml of culture (OD₆₀₀ = 1.0) was pelleted, rinsed, and resuspended in 2.5 ml Stainer Scholte (SS) medium [26] containing 25 µg/ml Cam and 20 µg/ml Str (SS+Cam+Str). Cultures were grown for 3 hr at 37°C to modulate bacteria to the Bvg⁺ phase and activate P_{tha} promoter expression. An aliquot of 500 µl from each culture was used for OD₆₀₀ measurement and cell number calculation. Phage particles were added to the rest of the culture at a multiplicity of infection of ~2.0. Following 1 hr incubation at 37°C for phage absorption, infected cells were pelleted and resuspended in 1 ml of fresh, prewarmed SS+Cam+Str media and incubated at 37°C for 3 hr post phage addition to allow completion of a single cycle of phage development. Progeny phages were harvested following chloroform extraction.

For phage production from lysogens, RB50 derivatives carrying appropriate prophages and donor plasmids were grown and modulated to the Bvg⁺ phase as in single-cycle lytic infections. Phage production was induced with 0.2 µg/ml mitomycin C for 3 hr at 37°C. Progeny phages were harvested by chloroform extraction.

BPP-1 phage tropism switching and PCR-based DGR homing assays

Phage tropism switching and DGR homing assays have been previously described [4].

Analysis of hairpin/cruciform formation in plasmid DNA *in vitro*

Plasmids containing the WT BPP-1 DGR target and the StMut mutation were isolated from *E. coli* DH5 α λpir cells using the QIAprep Spin miniprep kit (Qiagen). Plasmids were linearized by digestion with *Bgl*I as indicated. To analyze hairpin/cruciform structure formation in supercoiled or relaxed DNAs, 0.5 µg of supercoiled or linearized plasmids were treated with 10 units of T7 DNA endonuclease I (New England Biolabs, Ipswich, MA) for 40 minutes as in Miller *et al.* [11]. The reactions were terminated by phenol-chloroform-isoamyl alcohol (25:24:1) extraction and DNAs were precipitated with ethanol. T7 DNA endonuclease I cleavage sites were determined by primer extension with 5'-end ³²P-labeled primers using Vent (exo-) DNA polymerase (New

England Biolabs, Ipswich, MA) as in Miller *et al.* [11], except that 5% DMSO was added for GC-rich templates. Primer extension products were resolved on 6% polyacrylamide/8 M urea gels, alongside Sanger sequencing ladders generated with the same labeled primers and a plasmid template containing the WT target.

Targeting of a *Kan^R* gene by engineered BPP-1 phage DGRs

To target the *Kan^R* gene on a replicating phage, BPP-1 Δ 47R**Kan^S* phage particles were used for single-cycle lytic infection of RB50 cells transformed with appropriate donor plasmids, similar to phage production by single-cycle lytic infection described above. Progeny phages were titrated and ~10¹¹ pfu of different phages were added to 25 ml RB50 cells (OD₆₀₀ = 1.2) in SS+Str media for 8.0 hr to reestablish lysogens. Cells were pelleted and resuspended in 5 ml LB and serial dilutions were plated on LB+NA+Str and LB+NA+Str+Kan (50 µg/ml) to determine *Kan^R* gene targeting frequencies. Lysogen reestablishment efficiencies ranged from 60% to 100% based on PCR analysis of 10 colonies each picked on LB+NA+Str plates using phage specific primers. *Kan^R* targeting efficiency for each donor plasmid was determined as the ratio of colony forming units (cfu) on LB+NA+Str+Kan plates to those on LB+NA+Str, calibrated with the lysogen reestablishment efficiency for that sample.

To target the *Kan^R* gene on a prophage in the bacterial chromosome, RB50 cells lysogenized with phage BPP-1 Δ 47R**Kan^S* were transformed with appropriate donor plasmids. Starting cultures were grown overnight in LB+NA+Str+Cam as described above. An amount of cells equal to 1 ml of culture (OD₆₀₀ = 1.0) was pelleted, rinsed, and resuspended in 2.5 ml SS+Cam+Str and grown at 37°C for 6 hours. Serial dilutions were plated on LB+NA+Str and LB+NA+Str+Kan (50 µg/ml) to determine *Kan^R* gene targeting frequencies. *Kan^R* targeting efficiencies were determined as relative numbers of *Kan^R* cells as above. To target the *Kan^R* gene on a plasmid, the recipient plasmid pHGT-*Kan^S* and appropriate donors were transformed into RB50 cells and analyzed similarly. Tetracycline was added to 5.0 µg/ml for recipient plasmid maintenance.

Supporting Information

Figure S1 Alignment of BPP-1 DGR target deletion constructs showing deletion boundary sequences. (A) Alignment of 5' deletion constructs (Figure 1C) with the corresponding region of the WT sequence. The WT sequence extends from position -10 upstream of VR to VR position 134 (last nucleotide). The 5' end of VR and the (GC)₁₄ element for the WT sequence are marked. Sequences that replace the VR deletions in 5'Δ133 and 5'Δ153 are underlined in blue. The "inserted" sequences are significantly different from the original ones, although 5'Δ133 regains a C residue at position -1. (B) Alignment of 3' deletion constructs (Figure 1C) with the corresponding region of the WT sequence. The WT sequence extends from the 5' end of VR to the second codon of *avd*. The (GC)₁₄ element and the potential hairpin region for the WT sequence are marked. Sequences that replace the deletions in 3'Δ47, 3'Δ68, 3'Δ82 and 3'Δ103 are underlined in blue. The "inserted" sequences are significantly different from the original ones. Sequences downstream of the potential hairpin structure in 3'Δ54 and 3'Δ58 are shown in Figure 1C and are not aligned here. (PDF)

Figure S2 Alignment of homing products of recipient 5'Δ153 demonstrates cryptic 5' cDNA integration and adenine mutagenesis. (A) PCR detection strategy for homing products of recipient

5' Δ 153 and regions of the products aligned in (B) and (C). Primer annealing sites are indicated as small horizontal arrows. (B) Alignment of the homing products of recipient 5' Δ 153 from VR position 21 to the end of the TG2 tag shows cryptic cDNA integration sites and adenine mutagenesis. The recipient has a 5' deletion that includes the first 20 bp of VR. Cryptic integration sites are highlighted in pink. (C) Alignment of the transferred TG2 tag and its downstream VR sequence with the corresponding regions of the predicted homing product lacking adenine mutagenesis (TG2VR). (PDF)

Figure S3 Analysis of homing products of recipient StRev. (A) PCR detection strategy for homing products of recipient StRev and regions of the products aligned in (B) and (C). Primer annealing sites are indicated as small horizontal arrows. (B) Alignment of the homing products of recipient StRev from the first position of VR to the end of the TG2 tag with the corresponding region of the predicted homing product lacking adenine mutagenesis (VR5'end). Adenine mutagenesis is observed in 8/15 cloned homing products. (C) Alignment of homing products of recipient StRev from the beginning of TG2 to the start codon of *adv* with the corresponding regions of the predicted WT homing product lacking adenine mutagenesis (wtHP3'end). The hairpin region is underlined in red to show complementary changes. Adenine mutagenesis is observed in 5/20 cloned homing products. (PDF)

Figure S4 Sequence analysis of tropism switching products of phage BPP-1 Δ ATR with WT hairpin/cruciform structures. Sequences from the beginning of VR to the start codon of *adv* of five progeny phages with switched tropisms were aligned with the corresponding region of the predicted WT homing product lacking adenine mutagenesis (TR99VR). The hairpin region is underlined in red and adenine mutagenesis is observed in all five progeny phages with switched tropisms. (PDF)

Figure S5 Sequence analysis of tropism switching products of phage BPP-1 Δ ATR StMut. Sequences from the beginning of VR to the start codon of *adv* of five tropism-switched progeny phages of recipient StMut were aligned with the corresponding region of the predicted WT homing product lacking adenine mutagenesis (TR99VR). The hairpin region is underlined in red to show disruption of the structure. Adenine mutagenesis is observed in all five phage tropism switching products. (PDF)

Figure S6 Sequence analysis of tropism switching products of phage BPP-1 Δ ATR StRev. Sequences from the beginning of VR to the start codon of *adv* of five tropism-switched progeny phages of recipient StRev were aligned with the corresponding region of the predicted WT homing product lacking adenine mutagenesis (TR99VR). The hairpin region is underlined in red to show complementary changes. Adenine mutagenesis is observed in all five phage tropism switching products. (PDF)

Figure S7 Analysis of homing products of recipient VRInv. (A) PCR detection strategy for homing products of recipient VRInv and regions of the products aligned in (B) and (C). Primer annealing sites are indicated as small horizontal arrows. (B) Alignment of homing products of recipient VRInv from the 5' end of VR to the end of the TG2 tag with the corresponding region of the predicted homing product lacking adenine mutagenesis (VR5'end). Adenine mutagenesis is observed in 5/10 cloned

homing products. (C) Alignment of homing products of recipient VRInv from the beginning of TG2 to the end of VR with the corresponding region of the predicted WT homing product lacking adenine mutagenesis (VR3'end). Adenine mutagenesis is observed in 2/9 cloned homing products. (PDF)

Figure S8 Outline of marker coconversion assay with recipient phage BPP-1 Δ ATRSpl6. (A) PCR-based DGR homing assays with marked donor plasmids. Markers were introduced into plasmid pMX-TG1cAA [4]. The TR contains a 36 bp insert (TG1) at position 84. Grey and pink arrows represent TR and progeny VRs, respectively. Small horizontal arrows indicate primers used for homing assays: P7 and P8 are sense- and antisense-strand primers annealing upstream and downstream of VR, respectively; P9 and P10 (Table 1) are sense- and antisense-strand primers, respectively, that anneal to TG1. *Cam^R*, chloramphenicol resistance gene. (B) Schematic of coconversion experiments to determine 3' marker transfer boundaries. Single C to T markers downstream of the TG1 tag in donor TRs are indicated and the constructs have been previously described [4]. Markers (red T residues) are transferred to VR only if they are located between the TR positions corresponding to 3' and 5' cDNA integration sites in VR. (PDF)

Figure S9 Sequence analysis of replicating phage *Kan^R* targeting products with the pMX-Km1 donor. Sequences from the beginning of VR-*Kan^S* to the end of the hairpin structure were aligned with the corresponding region of the predicted *Kan^R* targeting product lacking adenine mutagenesis (KmHP). The targeting assay was carried out with BPP-1 Δ ATR**Kan^S* single-cycle lytic infection of RB50 cells transformed with donor plasmid pMX-Km1. Progeny phages were used to generate lysogens in RB50 cells, which were analyzed on plates with and without kanamycin to determine the efficiency of *Kan^R* targeting. *Kan^R* clones were sequenced to verify regeneration of full-length *Kan^R* genes. Adenine mutagenesis is observed in 7/10 clones. (PDF)

Figure S10 Sequence analysis of replicating phage *Kan^R* targeting products with the pMX-Km2 donor. Sequences from the beginning of VR-*Kan^S* to the end of the hairpin structure were aligned with the corresponding region of the predicted *Kan^R* targeting product lacking adenine mutagenesis (KmHP). The targeting assay was carried out with BPP-1 Δ ATR**Kan^S* single-cycle lytic infection of RB50 cells transformed with donor plasmid pMX-Km2. RB50 cells were lysogenized with progeny phages and subsequently analyzed on plates with and without kanamycin to determine the efficiency of *Kan^R* targeting. *Kan^R* clones were sequenced to verify regeneration of full-length *Kan^R* genes. Adenine mutagenesis is observed in 7/11 clones. (PDF)

Figure S11 Sequence analysis of prophage *Kan^R* targeting products with the pMX-Km1 donor. Sequences from the beginning of VR-*Kan^S* to the end of the hairpin structure were aligned with the corresponding region of the predicted *Kan^R* retargeting product lacking adenine mutagenesis (KmHP). The targeting assay was carried out in BPP-1 Δ ATR**Kan^S* lysogen cells transformed with donor plasmid pMX-Km1. Resulting cells were analyzed on plates with and without kanamycin to determine the efficiency of *Kan^R* targeting. *Kan^R* clones were then sequenced to verify regeneration of full-length *Kan^R* genes. Adenine mutagenesis is observed in 13/16 clones. (PDF)

Figure S12 Sequence analysis of prophage *Kan^R* targeting products with the pMX-Km2 donor. Sequences from the beginning of VR-*Kan^S* to the end of the hairpin structure were aligned with the corresponding region of the predicted *Kan^R* retargeting product lacking adenine mutagenesis (KmHP). The targeting assay was carried out in BPP-1Δ*ATR***Kan^S* lysogen cells transformed with donor plasmid pMX-Km2. Resulting cells were plated on plates with and without kanamycin to determine the efficiency of *Kan^R* targeting. *Kan^R* clones were sequenced to confirm regeneration of full-length *Kan^R* genes. Adenine mutagenesis is observed in 13/16 clones. (PDF)

Figure S13 Sequence analysis of plasmid *Kan^R* targeting products with the pMX-Km1 donor. Sequences from the beginning of VR-*Kan^S* to the end of the hairpin structure were aligned with the corresponding region of the predicted *Kan^R* targeting product lacking adenine mutagenesis (KmHP). Targeting assay was carried out in RB50 cells transformed with both recipient plasmid pHGT-*Kan^S* and donor plasmid pMX-Km1. Resulting cells were analyzed on plates with and without kanamycin to determine the efficiency of *Kan^R* targeting. *Kan^R* clones were sequenced to verify regeneration

of full-length *Kan^R* genes. Adenine mutagenesis is observed in 6/7 clones. (PDF)

Figure S14 Sequence analysis of plasmid *Kan^R* targeting products with the pMX-Km2 donor. Sequences from the beginning of VR-*Kan^S* to the end of the hairpin structure were aligned with the corresponding region of the predicted *Kan^R* retargeting product lacking adenine mutagenesis (KmHP). The targeting assay was carried out in RB50 cells transformed with both recipient plasmid pHGT-*Kan^S* and donor plasmid pMX-Km2. Resulting cells were plated on plates with and without kanamycin to determine the efficiency of *Kan^R* targeting. *Kan^R* clones were then sequenced to confirm regeneration of full-length *Kan^R* genes. Adenine mutagenesis is observed in 9/10 clones. (PDF)

Acknowledgments

We thank members of JFM laboratory and David W. Martin and other scientists at AvidBiotics for constructive input.

Author Contributions

Conceived and designed the experiments: HG JFM. Performed the experiments: HG LVT AWN EC SW SO VBL. Analyzed the data: HG JFM. Wrote the paper: HG JFM.

References

- Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, et al. (2004) Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature* 431: 476–481.
- Medhekar B, Miller JF (2007) Diversity-generating retroelements. *Curr Opin Microbiol* 10: 388–395.
- Liu M, Deora R, Doulatov SR, Gingery M, Eislering FA, et al. (2002) Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science* 295: 2091–2094.
- Guo H, Tse LV, Barbalat R, Sivaamnuaiaphorn S, Xu M, et al. (2008) Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Mol Cell* 31: 813–823.
- Zimmerly S, Guo H, Perlman PS, Lambowitz AM (1995) Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* 82: 545–554.
- 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.
- McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, et al. (2005) The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat Struct Mol Biol* 12: 886–892.
- Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, et al. (2008) Selective ligand recognition by a diversity-generating retroelement variable protein. *PLoS Biol* 6: e131. doi:10.1371/journal.pbio.0060131.
- Dai X, Greizerstein MB, Nadas-Chinni K, Rothman-Denes LB (1997) Supercoil-induced extrusion of a regulatory DNA hairpin. *Proc Natl Acad Sci U S A* 94: 2174–2179.
- Miller A, Dai X, Choi M, Glucksmann-Kuis MA, Rothman-Denes LB (1996) Single-stranded DNA-binding proteins as transcriptional activators. *Methods Enzymol* 274: 9–20.
- Lu M, Guo Q, Studier FW, Kallenbach NR (1991) Resolution of branched DNA substrates by T7 endonuclease I and its inhibition. *J Biol Chem* 266: 2531–2536.
- Mizushima T, Kataoka K, Ogata Y, Inoue R, Sekimizu K (1997) Increase in negative supercoiling of plasmid DNA in Escherichia coli exposed to cold shock. *Mol Microbiol* 23: 381–386.
- Witz G, Stasiak A (2010) DNA supercoiling and its role in DNA decatenation and unknotting. *Nucleic Acids Res* 38: 2119–2133.
- Siregar JJ, Miroshnikov K, Mobashery S (1995) Purification, characterization, and investigation of the mechanism of aminoglycoside 3'-phosphotransferase type Ia. *Biochemistry* 34: 12681–12688.
- Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, et al. (2004) Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol* 186: 1503–1517.
- Senior MM, Jones RA, Breslauer KJ (1988) Influence of loop residues on the relative stabilities of DNA hairpin structures. *Proc Natl Acad Sci U S A* 85: 6242–6246.
- Dai L, Zimmerly S (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30: 1091–1102.
- Robart AR, Seo W, Zimmerly S (2007) Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc Natl Acad Sci U S A* 104: 6620–6625.
- Granlund M, Michel F, Norgren M (2001) Mutually exclusive distribution of IS1548 and GBS11, an active group II intron identified in human isolates of group B streptococci. *J Bacteriol* 183: 2560–2569.
- Nishino T, Ishino Y, Morikawa K (2006) Structure-specific DNA nucleases: structural basis for 3D-scissors. *Curr Opin Struct Biol* 16: 60–67.
- Declais AC, Lilley DM (2008) New insight into the recognition of branched DNA structure by junction-resolving enzymes. *Curr Opin Struct Biol* 18: 86–95.
- Ichyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, et al. (2002) Retrotransposition of the L1LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* 46: 1259–1272.
- Zhong J, Lambowitz AM (2003) Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J* 22: 4555–4565.
- Lambowitz AM, Zimmerly S (2004) Mobile group II introns. *Annu Rev Genet* 38: 1–35.
- Stainer DW, Scholte MJ (1970) A simple chemically defined medium for the production of phase I Bordetella pertussis. *J Gen Microbiol* 63: 211–220.

Figure S1

A

┌→ VR

```
WT5'      ACGCCCAGCCCGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCCTCTCGGGTT 60
5'Δ105    ACGCCCAGCCCGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCCTCTCGGGTT 60
5'Δ133    GAAATGTTCCGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCCTCTCGGGTT 60
5'Δ153    CGCCACGAGCGCGTGGAAACAGAAATGTTTCGCGCCTGGAACGGCACGTCGCCTCTCGGGTT 60
           *      *      **      *      *      *      *****

WT        CTCGCGCTGCGCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCG 120
5'Δ105    CTCGCGCTGCGCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCG 120
5'Δ133    CTCGCGCTGCGCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCG 120
5'Δ153    CTCGCGCTGCGCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCG 120
           *****

WT        GCGTCTGTGACCACCTGATTCTTG 144
5'Δ105    GCGTCTGTGACCACCTGATTCTTG 144
5'Δ133    GCGTCTGTGACCACCTGATTCTTG 144
5'Δ153    GCGTCTGTGACCACCTGATTCTTG 144
           *****
```

B

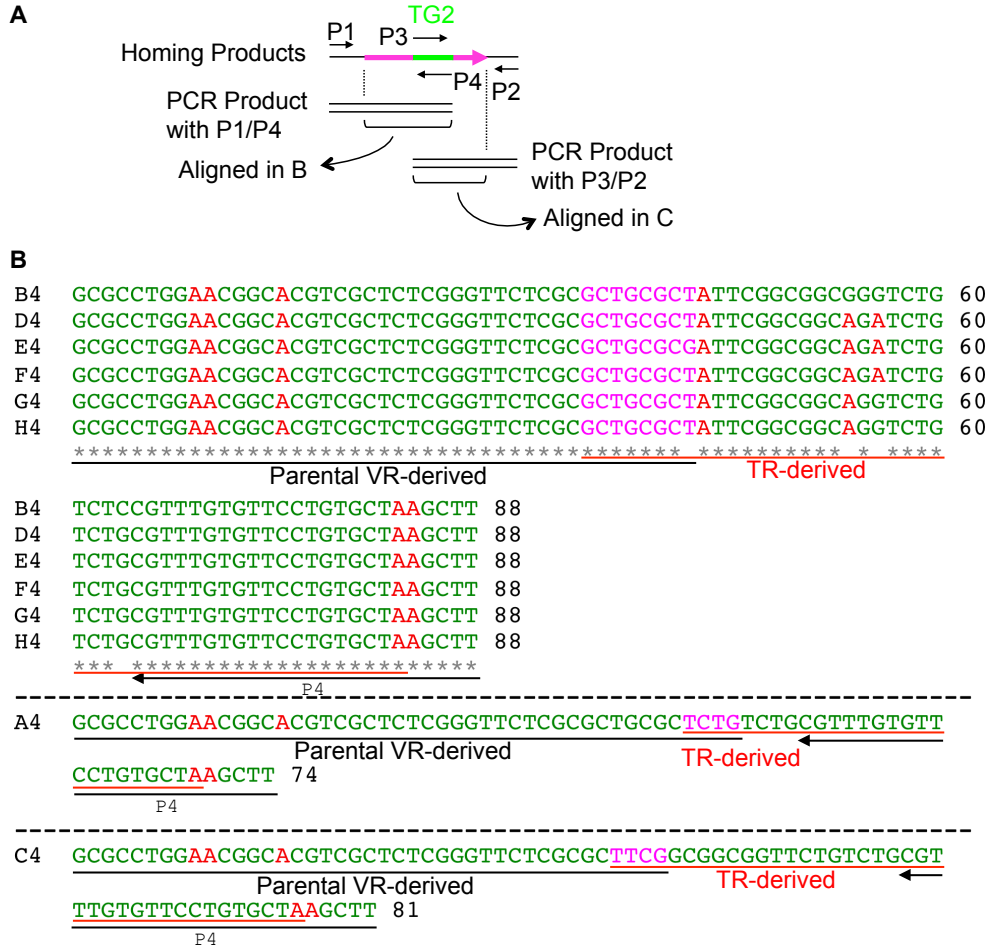
```
WT3'      CGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCTCTCGGGTTCTCGCGCTGC 60
3'Δ47     CGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCTCTCGGGTTCTCGCGCTGC 60
3'Δ68     CGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCTCTCGGGTTCTCGCGCTGC 60
3'Δ82     CGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCTCTCGGGTTCTCGCGCTGC 60
3'Δ103    CGCTGCTGCGCTATTCGGCGGCGCCTGGAACGGCACGTCGCTCTCGGGTTCTCGCGCTGC 60
           *****

WT3'      GCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCGCGCTCTGTGA 120
3'Δ47     GCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCGCGCTCTGTGA 120
3'Δ68     GCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCGCGCTCTGTGA 120
3'Δ82     GCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCGCGCTCTGTGA 120
3'Δ103    GCTCTGGTACAGCGGGCCGTCGTTCTCGTTCGCGTTCTTCGGGGCGCGCGCGCTCTGTGA 120
           *****

WT3'      CCACCTGATTCTTGAAGTAGCGGGCCGAAAAGGCCCGCCAAAGCAACCGATGGAA 175
3'Δ47     CCACCTGATTCTTGAAGTAGCGGGCCGAAAAGGCCCGCCAAAGCAACCGCTGCAG 175
3'Δ68     CCACCTGATTCTTGAAGTAGCGGGCCGACTGCAGCGGTTCTCGCCTCGTGGTTCG 175
3'Δ82     CCACCTGATTCTTGCATGAGCGGTTCTCGCCTCGTGGTTCGGGTCATGCGCAATG 175
3'Δ103    GGTTCCTCGCCTCGTGGTTCGGTATGCGCAATGGGCCGACACCCACAACCTTTT 175
           ** *      *

           Hairpin
```

Figure S2



C

```

TG2VR ---AGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCATCGGGGCGCGGGCGTCTGTGAC 57
A9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCGACCCATCGGGGCGCGGGCGTCTGTGGC 60
B9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCATCGGGGCGCGGGCGTCTGTGAC 60
C9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCTTCGGGGCGCGGGCGTCTGTGAC 60
D9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCATCGGGGCGCGGGCGTCTGTGAC 60
E9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCATCGGGGCGCGGGCGTCTGTGAC 60
F9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGTTAGCCATCGGGGCGCGGGCGTCTGTGAC 60
G9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCATCGGGGCGCGGGCGTCTGTGAC 60
H9    TCTAGATCTGTCTGCGTTTGTGTTCTGTGCTAGCCGTCGGGGCGCGGGCGTCTGTGAC 60

```

```

*****
P3  >
***** * * * *****
G/C

```

```

TG2VR CACCTGATTCTTG 70
A9    CACCTGATTCTTG 73
B9    CACCTGATTCTTG 73
C9    CACCTGATTCTTG 73
D9    CACCTGATTCTTG 73
E9    CACCTGATTCTTG 73
F9    CACCTGATTCTTG 73
G9    CACCTGATTCTTG 73
H9    CACCTGATTCTTG 73
*****

```

Figure S3



Figure S4

```
TR99VR CGCTGCTGCGCTATTCGGCGGCAACTGGAAACAACCGTCGAACTCGGGTTCTCGCGCTGC 60
wtTSP1 CGCTGCTGCGCTATTCGGCGGCGCCTGGAGCAACCGTCGAGCTCGGGTTCTCGCGCTGC 60
wtTSP2 CGCTGCTGCGCTATTCGGCGGCCTCTGGAGCAACCGTCGTACTCGGGTTCTCGCGCTGC 60
wtTSP3 CGCTGCTGCGCTATTCGGCGGCCTCTGGAGCAACCGTCGAACTCGGGTTCTCGCGCTGC 60
wtTSP4 CGCTGCTGCGCTGTTTCGGCGGCCTCTGGAGCAACCGTCGAACTCGGGTTCTCGCGCTGC 60
wtTSP5 CGCTGCTGCGCTATTCGGCGGCCTCTGGAGCAACCGTCGAACTCGGGTTCTCGCGCTGC 60
*****
VR
TR99VR GAACTGGAAACAACGGGCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA 120
wtTSP1 GTACTGGCACAGCGGGCCGTCGTACTCGTTCGCTTCTTCGGGCGCGCGGCGTCTGTGA 120
wtTSP2 GTACTGGAGCTACGGGCGTCGCTCTCGTTCGCTTCATCGGGCGCGCGGCGTCTGTGA 120
wtTSP3 GTACTGGTACAACGGGCGTCGAACTCGGCCGCTACTTCGGGCGCGCGGCGTCTGTGA 120
wtTSP4 GTACTGGAGCGCGGGCCGTCGTTCTCGTTCGCTTCTTCGGGCGCGCGGCGTCTGTGA 120
wtTSP5 GTACTGGAACTACGGGCGTCGTACTCGAACGCGTACATCGGGCGCGCGGCGTCTGTGA 120
*****
VR
TR99VR CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
wtTSP1 CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
wtTSP2 CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
wtTSP3 CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
wtTSP4 CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
wtTSP5 CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
*****
VR WT Hairpin
```

Figure S5

```

TR99VR      CGCTGCTGCGCTATTCGGCGGGCAACTGGAAACAACGTCGAACTCGGGTTCTCGCGCTGC 60
MutTSP6     CGCTGCTGCGCTATTCGGCGGGCAGCTGGAGCAACACGTCGAACTCGGGTTCTCGCGCTGC 60
MutTSP7     CGCTGCTGCGCTATTCGGCGGGCTCTGGAGCAACTCGTCGGCCTCGGGTTCTCGCGCTGC 60
MutTSP8     CGCTGCTGCGCTATTCGGCGGGCGCTGGAGCAACACGTCGAACTCGGGTTCTCGCGCTGC 60
MutTSP9     CGCTGCTGCGCTATTCGGCGGGCAGCTGGAGCAACACGTCGAACTCGGGTTCTCGCGCTGC 60
MutTSP10    CGCTGCTGCGCTATTCGGCGGGCGCTGGAGCAACACGTCGTCCTCGGGTTCTCGCGCTGC 60
*****
                                 VR
TR99VR      GAACTGGAAACAACGGGCCGTCGAACTCGAACCGCAACATCGGGGCGCGGGCGTCTGTGA 120
MutTSP6     GTA CTGGTACTACGGGCCGTCGTA CTGAGCGCGTACCTCGGGGCGCGGGCGTCTGTGA 120
MutTSP7     GTA CTGGGACTACGGGCCGTCGACCTCGTGC CGGTACA TCGGGGCGCGGGCGTCTGTGA 120
MutTSP8     GTA CTGGAGCTACGGGCCGTCGAGCTCGGGCGCGTACTTCGGGGCGCGGGCGTCTGTGA 120
MutTSP9     GTA CTGGAGCAGCGGGCCGTCGAACTCGTTCGCGTTCTTCGGGGCGCGGGCGTCTGTGA 120
MutTSP10    GGCCTGGAGCTACGGGCCGTCGGGCTCGGGCGCGTTCA TCGGGGCGCGGGCGTCTGTGA 120
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
                                 VR
TR99VR      CCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCGCCAAAGGCAACCGATG 172
MutTSP6     CCACCTGATTCTTGTAGTAGCGGGGCCGAAACCGGGGCCAAAGGCAACCGATG 172
MutTSP7     CCACCTGATTCTTGTAGTAGCGGGGCCGAAACCGGGGCCAAAGGCAACCGATG 172
MutTSP8     CCACCTGATTCTTGTAGTAGCGGGGCCGAAACCGGGGCCAAAGGCAACCGATG 172
MutTSP9     CCACCTGATTCTTGTAGTAGCGGGGCCGAAACCGGGGCCAAAGGCAACCGATG 172
MutTSP10    CCACCTGATTCTTGTAGTAGCGGGGCCGAAACCGGGGCCAAAGGCAACCGATG 172
*****
                                 VR
                                 Hairpin Region

```

Figure S6

```

TR99VR      CGCTGCTGCGCTATTCGGCGGGCAACTGGAAACAACGTCGAACTCGGGTTCTCGCGCTGC 60
RevTSP11    CGCTGCTGCGCTATTCGGCGGGCGCCTGGGCCAACACGTCGAACTCGGGTTCTCGCGCTGC 60
RevTSP12    CGCTGCTGCGCTATTCGGCGGGCGCCTGGAGCAACACGTCGTAACGTCGGTTCTCGCGCTGC 60
RevTSP13    CGCTGCTGCGCTATTCGGCGGGCTCCTGGAAACAACGTCGAACTCGGGTTCTCGCGCTGC 60
RevTSP14    CGCTGCTGCGCTATTCGGCGGGCTCCTGGAGCAACAACGTCGAGCTCGGGTTCTCGCGCTGC 60
RevTSP15    CGCTGCTGCGCTATTCGGCGGGCGCCTGGCTCCTACGTCGTTCTCGGGTTCTCGCGCTGC 60
*****
                                 VR
TR99VR      GAACTGGAAACAACGGGCGTTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA 120
RevTSP11    GTAAGTGGAGCTACGGGCGTTCGTCCTCGTACGCGTACATCGGGGCGCGGGCGTCTGTGA 120
RevTSP12    GTAAGTGGAAACAACGGGCGTTCGTTCTCGTTCGCGTTCTTCGGGGCGCGGGCGTCTGTGA 120
RevTSP13    GTAAGTGGGCGTTCGGGCGTTCGAACTCGTACGCGTACATCGGGGCGCGGGCGTCTGTGA 120
RevTSP14    GTAAGTGGAGCTACGGGCGTTCGAACTCGCACGCGTACATCGGGGCGCGGGCGTCTGTGA 120
RevTSP15    GTAAGTGGAGCTACGGGCGTTCGAACTCGCACGCGTACATCGGGGCGCGGGCGTCTGTGA 120
*****
                                 VR
TR99VR      CCACCTGATTCTTGTAGTAGGCCCCGGGAAAGGCCCGCCAAAGCAACCGATG 172
RevTSP11    CCACCTGATTCTTGTAGTAGGCCCCGGGAAACCGGGGCCAAAGCAACCGATG 172
RevTSP12    CCACCTGATTCTTGTAGTAGGCCCCGGGAAACCGGGGCCAAAGCAACCGATG 172
RevTSP13    CCACCTGATTCTTGTAGTAGGCCCCGGGAAACCGGGGCCAAAGCAACCGATG 172
RevTSP14    CCACCTGATTCTTGTAGTAGGCCCCGGGAAACCGGGGCCAAAGCAACCGATG 172
RevTSP15    CCACCTGATTCTTGTAGTAGGCCCCGGGAAACCGGGGCCAAAGCAACCGATG 172
*****
                                 VR
                                 Hairpin StemRev

```

Figure S7

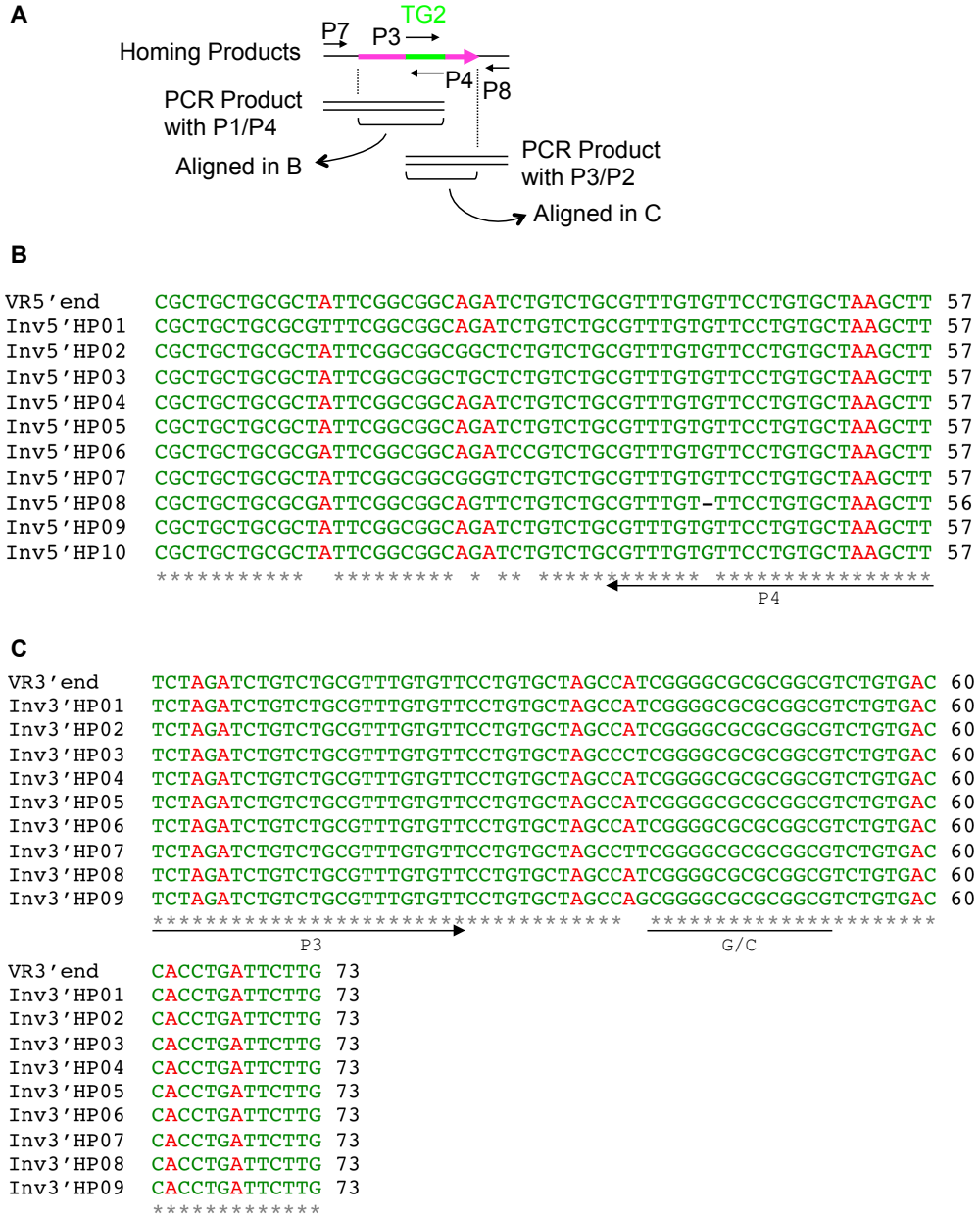


Figure S8

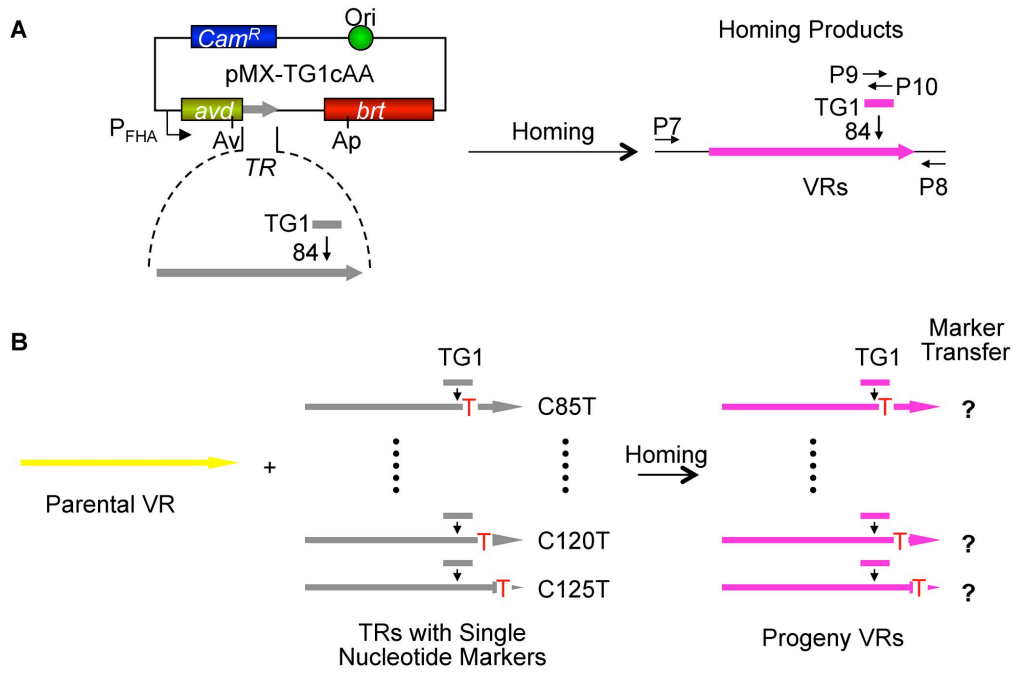


Figure S9

```

KmHP      CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP1-01    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTGGCCGTCGGGGCGCG 60
RP1-02    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP1-03    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
RP1-04    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTGGCCATCGGGGCGCG 60
RP1-05    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP1-06    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP1-07    CGCTTGCAGTTTCATTTGGTGCTCGATGAGTTTTTCTAGTAAGCTCGCCATCGGGGCGCG 60
RP1-08    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTGGCCATCGGGGCGCG 60
RP1-09    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTGGCCATCGGGGCGCG 60
RP1-10    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAATAAGCTAGCCGTCGGGGCGCG 60
          *****
          Regenerated KanR 3' end
          ***** GC

KmHP      CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-01    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-02    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-03    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-04    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-05    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-06    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-07    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-08    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-09    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP1-10    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
          *****
          WT Hairpin
    
```

Figure S10

```

KmHP      CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-01    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-02    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGGGCTAGCCATCGGGGCGCG 60
RP2-03    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-04    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-05    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-06    CGCTTGCAGTTTCATTTGGTGCTCGATGAGTTTTTCTAATAAGCTAGGCCTCGGGGCGCG 60
RP2-07    CGCTTGCAGTTTCATTTGCTGCTCGATGAGTTTTTCTAATGGGCTGGCCGTCGGGGCGCG 60
RP2-08    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-09    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
RP2-10    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTGGCCATCGGGGCGCG 60
RP2-11    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
          *****
          Regenerated KanR 3' end          GC

KmHP      CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-01    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-02    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-03    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-04    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-05    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-06    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-07    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-08    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-09    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-10    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
RP2-11    CGGCGTCTGTGACCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCCGC 110
          *****
          WT Hairpin
    
```

Figure S11

```

KmHP      CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-01    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTGAGCTAGCCATCGGGGCGCG 60
LY1-02    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-03    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATGGGCTAGCCATCGGGGCGCG 60
LY1-04    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
LY1-05    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATAAGCTAGCCTTCGGGGCGCG 60
LY1-06    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
LY1-07    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
LY1-08    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAGTAAGCTAGCCCTCGGGGCGCG 60
LY1-09    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
LY1-10    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-11    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-12    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-13    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
LY1-14    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-15    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY1-16    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCGTCGGGGCGCG 60
          *****
          Regenerated KanR 3' end          *          *****          GC

KmHP      CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-01    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-02    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-03    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-04    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-05    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-06    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-07    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-08    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-09    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-10    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-11    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-12    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-13    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-14    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-15    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
LY1-16    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
          *****
          WT Hairpin
    
```

Figure S12

```

KmHP      CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY2-01    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTGTCTAATAGGCTAGCCATCGGGGCGCG 60
LY2-02    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
LY2-03    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY2-04    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
LY2-05    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
LY2-06    CGCTTGCAGTTTCATTTGATGCTCGCTGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
LY2-07    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY2-08    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGGGCTAGCCATCGGGGCGCG 60
LY2-09    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTGGGCTGGCCATCGGGGCGCG 60
LY2-10    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY2-11    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTGATAAGCTAGCCATCGGGGCGCG 60
LY2-12    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
LY2-13    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
LY2-14    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTAGCCCTCGGGGCGCG 60
LY2-15    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
LY2-16    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
          *****
          Regenerated KanR 3' end      *   ***   **   *****
          G/C

KmHP      CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-01    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-02    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-03    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-04    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-05    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-06    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-07    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-08    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-09    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-10    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-11    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-12    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-13    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-14    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-15    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
LY2-16    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCGGAAAGGCCCCGC 110
          *****
          WT Hairpin
    
```

Figure S13

```

KmHP      CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
PL1-01    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
PL1-02    CGCTTGCAGTTTCATTTGATGCTCGCTGAGTTTTTCTAGTTTGCTAGCCCTCGGGGCGCG 60
PL1-03    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGGGCTAGCCATCGGGGCGCG 60
PL1-04    CGCTTGCAGTTTCATTTGGTGCTCGATGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
PL1-05    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATCTGCTAGCCATCGGGGCGCG 60
PL1-06    CGCTTGCAGTTCCATTTGATGCTCGGTGAGTTTTTCTAATAAGCTGGCCATCGGGGCGCG 60
PL1-07    CGCTTGCAGTTTCATTTGATGCTCGTTGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
          *****  *****  *****  *****  *****  *  ***  **  *****
          Regenerated KanR 3' end                                     GC
KmHP      CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-01    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-02    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-03    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-04    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-05    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-06    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL1-07    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
          *****
          WT Hairpin
    
```

Figure S14

```

KmHP      CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
PL2-01    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
PL2-02    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGGGCTAGCCATCGGGGCGCG 60
PL2-03    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTAGCCATCGGGGCGCG 60
PL2-04    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGCTGGCCATCGGGGCGCG 60
PL2-05    CGCTTGCAGTTTCATTTGTTGCTCGATGAGTTTTTCTAATAAGCTGGCCGTCGGGGCGCG 60
PL2-06    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTGATGGGCTGGCCATCGGGGCGCG 60
PL2-07    CGCTTGCAGTTTCATTTGATGCTCGGTGAGTTTTTCTAGTAAGCTAGCCATCGGGGCGCG 60
PL2-08    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATGAGCTAGCCATCGGGGCGCG 60
PL2-09    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAGTAAGGTAGCCGTCGGGGCGCG 60
PL2-10    CGCTTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATAAGCTGGCCATCGGGGCGCG 60
          *****
          Regenerated KanR 3' end          * * * * * GC
KmHP      CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-01    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-02    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-03    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-04    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-05    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-06    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-07    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-08    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-09    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
PL2-10    CGGCGTCTGTGACCACCTGATTCTTGTAGTAGCGGGGCCGAAAGGCCCCGC 110
          *****
          WT Hairpin

```

CHAPTER 3. *Cis*-acting DNA Structural Elements Guide Targeted Mutagenesis by Diversity Generating Retroelements

ABSTRACT

Diversity-generating retroelements (DGRs) are a family of retroelements capable of generating vast amounts of nucleotide variability within defined protein-encoding DNA sequences. Diversification of target sequences is site specific and occurs through a distinct reverse transcriptase-mediated process called mutagenic homing. DGRs are widely distributed in nature and have been identified in plasmids, bacteriophage, and bacterial and archaeal genomes. We have demonstrated that mutagenic homing requires both specific sequence and structural elements, including target recognition sequences that include a DNA stem-loop/cruciform structure. Stem-loops have been identified in most DGRs and while the length and sequence of the stem varies, the nucleotide composition of the loop is conserved among these structural elements. Using the *Bordetella* BPP-1 DGR, we demonstrate that conserved loop residues contribute to stem-loop/cruciform formation and stability, and are thus critical for DGR mutagenic homing. Additionally, our results indicate the polarity of the stem-loop is required for structure formation and the orientation of the loop nucleotide sequence determines target site recognition during mutagenic homing. Analysis of DGR stem-loops from disparate species indicates that these conserved elements are functionally interchangeable and fundamental to target site recognition.

INTRODUCTION

Diversity-generating retroelements (DGRs) are a family of retroelements capable of accelerating evolution of adaptive traits by generating nucleotide variability within protein-encoding DNA sequences [1-3]. While discovered in a bacteriophage (BPP-1) that infects the mammalian pathogen *Bordetella bronchiseptica* [1], DGRs have since been identified within members of most bacterial genera, within archaeal and their viruses [4]. The breadth of organisms which contain DGRs, along with their divergent lifestyles, indicate that they represent a conserved prokaryotic system for targeted protein evolution.

DGRs encode a reverse transcriptase (RT), an accessory protein (*avd*), a template repeat (TR), and a variable repeat (VR) which are required for diversification of target protein (TP) encoding gene(s). DGRs diversify DNA sequences through a process called mutagenic homing that introduces nucleotide substitutions into VRs of target genes (Figure 1). Sites of nucleotide substitutions in VR correspond to adenine nucleotides in TR, which is transcribed as an RNA intermediate and serves as a template for DGR-RT dependent cDNA synthesis. During reverse transcription TR-adenine residues are replaced with any DNA base, which results in a mutagenized cDNA that displaces the parental VR [5]. In the BPP-1 DGR the target gene, *mtd* (major tropism determinant), encodes the phage tail fiber protein responsible for binding to host-cell receptors [1, 6-8]. Thus DGR mutagenic homing generates BPP-1 Mtd variants that recognize new *B. bronchiseptica* cell surface molecules as ligands for infection.

While the introduction of nucleotide substitutions into target genes is required for DGR protein evolution, mutagenic homing must be constrained to specific sequences to

maintain genome integrity and avoid host loss of fitness. Previously we demonstrated that DGR target site recognition requires both nucleotide recognition sequences and structural elements [5, 9]. In the BPP-1 DGR these requirements are fulfilled by the IMH (initiation of mutagenic homing), an element composed of a 14 base pair G/C stretch, which is followed by a 21 bp sequence, and an inverted repeat that under physiological levels of negative supercoiling forms a stem-loop or double-stranded cruciform structure (Figure 1) [9]. Comparative bioinformatics revealed analogous stem-loops in the vast majority of DGRs, suggesting their conserved role in target site recognition. DNA and RNA stem-loop structures are common features of prokaryotic and eukaryotic genomes and play a mechanistic role in many biological processes including DNA replication, transcription regulation, and DNA repair [10-16]. DGR stem-loops contain a 3-4nt loop with the consensus sequence 5'GNA3' or 5'GRNA3' (where N=A, G, C or T and R=A, G). Stem-loops with GNA trinucleotide or GNRA tetranucleotide loops form unusually stable structures [10] due to the ability of the first and last residue of these loops to form a non-canonical sheared G•A base pairing. This type of base pairing has been reported to provide increased stem-loop structure stability compared to canonical Watson-Crick base interaction(s) by allowing extensive loop base pair stacking interactions and base stacking between the G•A base pair and residues flanking the loop [10-13, 17-21]. This suggests sheared base stacking may be important in DGR stem-loop formation and stability.

During DGR mutagenic homing, formation of a DNA stem-loop is necessary for efficient recognition and diversification of target genes [9]. However, while necessary the precise role stem-loops play in mutagenic homing has not been determined. Using

the model system BPP-1 we took a genetic approach to investigate how the formation, structure, and stability of DGR stem-loops influences target site recognition. Specifically, we investigated how loop nucleotide composition contributes to stem-loop formation and stability. In addition to influencing structure stability, orientation of the loop nucleotide sequence also plays a critical role in target site recognition. As stem-loops are a conserved feature of most DGRs, the ability of similar elements from disparate species to support mutagenic homing was analyzed and were found to be functionally interchangeable. Therefore, based on our examination of how these elements determine which nucleotide sequences are targeted for mutagenesis we propose a set of fundamental principles for target site recognition during DGR mutagenic homing.

RESULTS

Nucleotide composition of the tetranucleotide loop is critical for structure formation

The contribution of DNA stem-loops to mutagenic homing were investigated using a phage tropism switching assay that quantitatively assess the relative frequency of DGR activity because it requires adenine mutagenesis during retrohoming [1, 2, 5, 9]. For these assays, wild type (WT) or mutant stem-loops were introduced into BPP-1 lysogens where *avd*, TR and *brt* (ΔATR) had been deleted. These BPP-1 ΔATR lysogens were complemented with a donor plasmid (pMX1) expressing BPP-1 DGR *avd*, TR, and *brt* from a regulated promoter [9]. As a negative control, BPP-1 ΔATR lysogens were complemented with pMX1SMAA, a construct expressing an inactive bRT that is incapable of catalyzing mutagenic homing and thus these phage cannot switch tropisms.

The BPP-1 stem-loop is composed of a 8 bp stem with a 4 nt (GAAA) loop and we investigated if the G•A closing base pair is critical to target site recognition by replacing the first and fourth residues (GAAA) with their complementary nucleotides (SLM2A). As shown in Figure 2A, substitution of the G•A closing base pair resulted in a 10^5 fold reduction in tropism switching. Conversely, substitution of the second and third adenine residues with T (SLM2B), C (SLMGCCA), or G (SLMGGGA) had no detectable effect on tropism switching. The Mtd VR regions in WT, SLM2A, and SLM2B phage were sequenced to verify tropism switching was due to adenine mutagenesis (Figures S1-S3). These nucleotide substitutions could suppress mutagenic homing because of altered structure formation and therefore we analyzed the ability of these mutants to

form stem-loop or double-stranded cruciform structures *in vitro*. Supercoiled plasmids carrying WT or mutant BPP-1 target sequences were treated with T7 DNA endonuclease I, an enzyme that recognizes DNA four-way junctions and has been used to probe hairpin or cruciform structure formation, followed by primer extension using radiolabeled primers to detect strand cleavage [9, 11, 22]. Major and minor endonuclease-dependent cleavage sites were observed on both strands within WT and SLM2B BPP-1 stem-loop sequences, while cleavage was significantly reduced in the SLM2A loop mutant (Figure 2B). The reduced cleavage of SLM2A loop mutants suggested that stem-loops were less stable, so we probed their thermal stability using ultra-violet (UV) absorption spectroscopy [23, 24]. Consistent with previous reports [10-13, 17-19, 21], the GAAA loop appears to provide WT stem-loops with an unusual stability ($t_m = 89.3^\circ\text{C} \pm 0.7$) as indicated by a melting temperature greater than 82°C [21]. A similar melting temperature ($t_m = 91.0^\circ\text{C} \pm 1.1$) was observed for SLM2B (GTTA) loop mutant which retains the G•A closing base pair. As predicted, disruption of the loop G•A closing base pair reduced thermal stability in the SLM2A (CAAT) mutant ($t_m = 81.7^\circ\text{C} \pm 1.4$). These results highlight the importance of the loop closing base pair interaction in DGR stem-loop formation and stability.

To address if a non-canonical sheared G•A base pairing was required for mutagenic homing, we generated stem-loop mutants where the closing base pair was disrupted by substituting the fourth loop residue (GAAA) with G, T, or C. As shown in Figure 3A, substitution with T (SLMGT) or G (SLMGG) significantly decreased tropism switching while substitution with C (SLMGC) resulted in a tropism switching frequency

comparable to WT (Figure S4). We then determined if these substitutions affected DNA structure formation and found SLMGG, SLMGT, and SLMGC stem-loops were all susceptible to T7 endonuclease I cleavage (Figure 3B). Interestingly, this indicates that certain nucleotide substitutions supported stem-loop formation but not mutagenic homing (see discussion). Cumulatively, we identified key residues which contribute to stem-loop formation and target site recognition, and determined the loop nucleotide sequence can vary as long as closing base pair interaction sufficient for a stable stem-loop structure is maintained.

Stem-loop orientation determines sequence recognition during DGR mutagenic homing

While necessary for target site recognition during mutagenic homing, stem-loops can form on both DNA strands and likely form simultaneously to generate a structurally symmetrical cruciform [9]. However, nucleotide diversification is only observed within sequences upstream (relative to the VR) of the stem-loop/cruciform which suggests loop sequence orientation may contribute to the directionality of mutagenic homing. We determined the ability of mutant stem-loop sequences to serve as recipients for DGR retrohoming using an *in vitro* PCR-based assay as it allows for VR and hairpin modifications which inhibit Mtd function *in vivo* [5, 9]. BPP-1 Δ ATR lysogens were complemented with a derivative of the donor plasmid pMX1, where an invariant 30 bp DNA tag was inserted into TR (pMX- Δ TR23-96), and mutagenic retrohoming resulting in transfer of this to VR was detected by PCR (Figure 4B). BPP-1 Δ ATR prophage containing a VR with wt stem-loop (WTVR), inverted VR/stem-loop (VRInv) [9], or a

derivative of VRInv where the stem-loop was reverted back to its original orientation (VRInvWTSL) were tested for their competence in target site recognition (Figure 4A). As shown in Figure 4C, DGR homing products were detected in WTVR and VRInv stem-loops, with sequencing of homing products verifying adenine mutagenesis (Figure S5), but homing products were not detected in VRInvWTSL. This indicates that while they can serve as recognition elements on either DNA strand, the polarity of DGR stem-loops/cruciform structure and sequence determine which nucleotide sequences are subjected to mutagenic retrohoming.

We then investigated if loop sequence orientation affected mutagenic homing *in vivo* by generating prophage that contained either the reversed sequence of the BPP-1 stem-loop (SLMAG), the inverted loop nucleotide sequence (SLMTC), or SLMTC with the loop sequence reversed (SLM4). As shown in Figure 5A, a significant decrease in tropism switching frequency was observed for all three constructs when compared to WT. These mutants were then subjected to T7 endonuclease I to determine if the lack of tropism switching was due to absence of stem-loop formation. Endonuclease cleavage was not detectable in the SLMAG mutant which suggests switching the directionality of the stem-loop sequence had profound effects on structure formation. However, cleavage sites were observed on both strands for SLMTC and SLM4 mutants and this demonstrates stem-loop structure formation occurs regardless of loop sequence orientation (Figure 5B). The observation that SLMTC only varies from WT in loop sequence but cannot support mutagenic homing is consistent with our *in vitro*

experiments and confirms the importance of stem-loop orientation during target site recognition.

DGR stem-loops are functionally interchangeable between species

Comparative bioinformatics identified stem-loops as a conserved feature of DGRs, including those recently identified in Archaea [4], which is consistent with their fundamental role as recognition elements. While target site recognition has been mostly studied using BPP-1, the importance of stem-loops in mutagenic homing has also been demonstrated in chromosomally encoded *Legionella pneumophila* DGRs [25].

Therefore, to address if DGR target site recognition functions by conserved structural and sequence constraints, we used BPP-1 to investigate if stem-loops from disparate species are functionally interchangeable. BPP-1 Δ ATR prophage carrying stem-loop sequences from *L. pneumophila* D5591, *L. pneumophila* Corby, *Bacteroides fragilis* 638R, *Bacteroides ovatus* ATCC 8483 or the DUSEL4 *Nanoarchaeaota* (Figure 6A) were assessed for their competence in supporting DGR mutagenic homing using tropism switch assays. As show in Figure 6B, tropism switching could not be detected in BPP-1 containing stem-loop sequence from *L. pneumophila* D5591, *L. pneumophila* Corby, *B. fragilis* 638R, and DUSEL4 *Nanoarchaeaota*. However, substituting the BPP-1 stem-loop with the *B. ovatus* ATCC 8483 sequence resulted in detectable tropism switching, although with decreased efficiency (Figure 6B). While all tested stem-loops contained the G•A closing base pair, the *B. ovatus* is the only element with a 4nt loop and we hypothesize the decreased levels of tropism switching result from variance in G+C content of the stem.

Since the size of the loop is essential for stem-loop function in target recognition, we addressed whether tropism switching in BPP-1 containing the *L. pneumophila* D5591 stem-loop could be rescued by increasing the size of the loop. Stem-loops composed of a 4nt loop were created by inserting a single nucleotide between the second and third residue of the loop (GCXA, where X=A, G, C or T) and their ability to support mutagenic homing was assessed (Figure 6C). In contrast to BPP-1 phage carrying the three nucleotide D5591 stem-loop, tropism switching was observed for phage containing stem-loops with loop nucleotide sequence GCAA (LPGCAA) or GCCA (LPGCCA) but not with GCGA (LPGCGA) and GCTA (LPGCTA) suggesting that loop nucleotide composition contributes to stem-loop function in target site recognition or these nucleotide insertions resulted in unanticipated interactions. Taken together, these results suggest that DGR stem-loops are fundamental to target site recognition and variations in sequence or structure represent host specific coevolution.

DISCUSSION

Of all DNA mutagenic systems found in nature, DGRs can introduce the greatest number of substitutions into protein encoding genes and here we explored how hosts maximize this potential by confining mutagenic homing to well-defined nucleotide domains. Using BPP-1 as a model system, the structure and sequence composition of DGR stem-loops were found to be essential to target site recognition during mutagenic homing. Most DGRs stem-loops contain the loop sequence GNNA and we demonstrated both conserved residues are required for structure formation, increased thermal stability, and mutagenic homing dependent tropism switching. Substitution of the fourth loop residue (GAAA) with T or G allowed stem-loop formation but did not support mutagenic homing. Given the importance of the G•A interaction to thermal stability, nucleotide substitutions which disrupt the closing base pair interaction likely retain stem-loop formation but with dampened stability and increased disassociation rate. Alternatively, substitution of the fourth loop residue could favor interactions between the first and third residue of the loop (GAAA) resulting in stem-loop structures with a 3nt loop [10-13, 17, 20, 26]. The hypothesis that mutants SLMGT and SLMGG form 3nt stem-loops which cannot support mutagenic homing is supported by previous reports where alteration in the number of loop residues drastically reduced levels of tropism switching [9]. While most substitutions inhibited DGR activity, we found loops with a GAAC sequence formed stem-loops and supported mutagenic homing, which is likely attributed to the ability of G and C residues to form a sheared-like G•C base pairing [13, 27]. While a G•C interaction is sufficient for structure formation and mutagenic homing, the observation that DGR stem-loops are predominately composed

of GNRA or GNA residues [10-13,17-19, 20, 21] suggests restricting loop closing base pairs to G•A reflects selective pressures beyond target site recognition.

DGR stem-loop formation likely occurs on both the template and non-template strand simultaneously to form a structurally symmetrical DNA cruciform. However, our *in vitro* and *in vivo* DGR homing assays demonstrated inversion of just the stem-loop sequence (relative to VR) did not affect stem-loop formation but eliminated mutagenic homing. This indicates stem-loop/cruciform structures are functionally non-symmetrical and the directionality of mutagenic homing is likely determined by the loop nucleotide sequence. The polar binding of DNA/RNA stem-loops by proteins has also been observed as integral to the preferential recognition of nucleotide strands during cellular processes like replication, recombination, and transcription [11, 15, 29]. Thus we hypothesize DGR stem-loop sequence are recognized by the AVD-RT complex or other host factors to facilitate DNA processing events, like nicking of the IMH or cDNA priming, required during mutagenic homing.

Whether found in a phage, bacterial, or archaeal host, we hypothesize DGR mutagenic homing target site recognition would function through a conserved mechanism. Therefore we investigated if stem-loops from bacterial and archaeal DGRs were functionally interchangeable with a phage element. Our analysis found stem-loops from *L. pneumophila* and *B. ovatus* could support mutagenic homing in BPP-1, albeit with lower efficiency. Furthermore, the observation that modification of the loop sequence of a bacterial stem-loop to be more “phage like” resulted in increased tropisms switching confirms a coevolutionary relationship, with individual DGR components optimized towards host needs. Based on these and previous analyses [9],

we propose the function of DGR stem-loops in target site recognition is determined by three fundamental principles: 1) size and nucleotide composition of the loop, 2) base pair composition of the stem, and 3) orientation of the stem-loop/cruciform structure (Figure 7). These principles will guide future work on refining our understanding of how target site recognition occurs during DGR mutagenic homing.

Figure legends:

Figure 1. Proposed mechanism for DGR-mediated mutagenic homing.

The BPP-1 DGR encodes a target gene (*mtd*), an accessory protein (*avd*), a dedicated reverse transcriptase (*bRT*), a variable repeat (VR) and an invariant template repeat (TR). DGR-mediated sequence diversification results from the introduction of nucleotide substitutions in VR which is located at the 3' end of *mtd*. Sites of nucleotide substitutions in VR (black arrow) correspond to adenine nucleotides in TR (black arrow), which is invariant and serves as a template to derive an RNA intermediate. During reverse transcription, adenine nucleotides are replaced with random nucleotides and the synthesized cDNA (yellow arrow) displaces the parental VR, a process termed mutagenic homing. Target recognition during mutagenic homing requires an IMH element (initiation of mutagenic homing) which is located at the 3' end of VR (purple box). Components of the IMH are depicted in the expanded view, and include a 14 bp G/C-rich region followed by a 21bp DNA segment and an inverted repeat that forms a stem-loop structure (pink) composed of a 8bp stem and a 4nt loop [9].

Figure 2. The nucleotide composition of the loop influences stem-loop/ cruciform formation and is critical for DGR mutagenic homing.

(a) Loop closing residues, as opposed to residues at the center of the loop are essential for phage tropism switching. The graph illustrates phage tropism switching frequencies (TSF) for BPP-1 Δ ATR carrying a wild-type stem-loop (WT) or loop mutants (SLM2A, SLM2B, SLMGCCA or SLMGGGA). Stem-loop sequences are shown (top) and nucleotide modifications to the loop are indicated in red. Donor RT indicates pMX-1

plasmid expressing a functional RT (+) or an enzymatically inactive RT (-). No phage plaques were detected for RT-deficient donors and the TSF shown represent limits of detection. The bars represent average TSF values from three independent experiments, and error bars represent \pm SEM. Asterisks indicate P values comparing mutants to WT in Student's T tests. * $P < 0.05$. (b) The loop G•A base pairing contributes to stem-loop/cruciform structure formation. Primer extension analysis of T7 endonuclease I cleavage of supercoiled plasmid DNA carrying a WT or modified stem-loop sequences (StMut, SLM2A, SLM2B). Stem-loop sequences are shown above the gels and nucleotide modifications are shown in red. Top strand cleavage (left panel) and bottom strand cleavage (right panel) is shown. Plasmid DNA carrying WT stem-loop sequence was either untreated (-) or treated (+) with T7 endonuclease I (T7) followed by primer extension. Sequence ladder on the left side of each panel and primers extension termination sites shown below the gels correspond to WT plasmid DNA. Black arrows indicate major cleavage sites and blue arrows indicate minor cleavage sites. Atd5' and VR3' are primers used in primer extension assays.

Figure 3. Base pairing within the loop is required for DGR mutagenic homing.

(a) The loop closing base pair is essential for tropism switching. Tropism switching frequencies observed for BPP-1 Δ ATR with a WT or loop mutants (SLMGG, SLMGT or SLMGC) are shown. Stem-loop sequences are shown (top) and nucleotide modifications to the loop are indicated in red. Donor RT indicates pMX-1 plasmid expressing a functional RT (+) or an enzymatically inactive RT (-). No phage plaques were detected for RT-deficient donors and the TSF shown represent limits of detection.

The bars represent average TSF values from three independent experiments, and error bars reflect \pm SEM. Asterisks indicate P values comparing mutants to WT in Student's T tests. *P<0.05. (b) Effect of loop closing base pair substitutions on stem-loop/cruciform structure formation. Primer extension analysis of T7 endonuclease I cleavage of plasmid DNA carrying WT or stem-loop mutants (StMut, SLMGG, SLMGT, SLMGC) was performed as stated above. Stem-loop sequences are shown above the gels, and nucleotide modifications are indicated in red. Top strand cleavage sites (left panel) and bottom strand cleavage sites (right panel) are shown. Thick black arrows indicate major cleavage sites and blue arrows indicate minor cleavage sites.

Figure 4. Stem-loop orientation is critical for sequence recognition during BPP-1 DGR retrohoming. (a) Diagram illustrates DGR components from BPP-1 Δ ATR with WT (WTVR) or modified (VRINV, VRINVWTSL) target sequences. In BPP-1 Δ ATRVRInv, a segment of *mtd* from position -133 upstream of VR to position +82 downstream of VR was inverted [9]. BPP1 Δ ATRVRInvwtSL is a derivative of VRInv where the stem-loop was inverted back to its original orientation. (b) The diagram shows primers used in DGR homing assays. P3 and P4 primers anneal to the 30 bp PCR tag (shown in green) and are indicated by small horizontal arrows. P7 and P8 primers anneal upstream and downstream of VR, respectively. (c) Stem-loop/cruciform orientation (relative to VR) is critical for target sequence recognition during mutagenic retrohoming. Plasmid pMX- Δ TR23-96 expressing a functional RT (+) or an inactive RT (-) was used as donor for the indicated recipients. Products from PCR-homing assays with primer pairs (P7+P4,

P3+P8, P7+P8) are shown, and the transfer of the PCR tag from donor plasmid TR to the VRINV recipient was verified by sequence analysis of homing products (Sup. fig. 5).

Figure 5. Stem-loop orientation determines sequence recognition during BPP-1 DGR mutagenic homing.

(a) Tropism switching frequencies observed for BPP-1 Δ ATR with a WT stem-loop or loop mutants (SLM4, SLMAG or SLMTC) are shown. Stem-loop sequences are shown (top) and loop nucleotide modifications are indicated in red. For SLM4 the nucleotide sequence of the loop was changed to the opposite residues. In SLMAG, the stem-loop sequence was reversed while in SLMTC the stem-loop sequence was inversed. Donor RT indicates pMX-1 plasmid expressing a functional RT (+) or an enzymatically inactive RT (-). No phage plaques were detected for RT-deficient donors and the TSF shown represent limits of detection. Bars represent average TSF values from three independent experiments, and error bars represent \pm SEM. Asterisks indicate P values comparing mutants to WT in Student's T tests. *P<0.05. (b) Effects of sequence orientation on stem-loop/cruciform structure formation. Primer extension analysis of T7 endonuclease I cleavage of supercoiled plasmid DNA carrying WT or stem-loop mutants (SLM4, SLMAG, SLMTC) was performed as stated above. Stem-loop sequences are shown above the gels, and nucleotide modifications are indicated in red. Top strand cleavage sites (left panel) and bottom strand cleavage sites (right panel) are shown. Black arrows designate major cleavage sites and blue arrows indicate minor cleavage sites.

Figure 6. DGR stem-loop structures are functionally interchangeable between species. (a) Schematics of DGR cassettes from bacterial (*L. pneumophila* D5591, *L. pneumophila* Corby, *B. fragilis* 638R and *B. ovatus* ATCC 8483) and archaeal (DUSEL4 *Nanoarchaeota* OTU2) spp. Inverted repeats that have the potential to form stem-loop/cruciform structures were identified downstream of VR as shown. Nucleotide sequence of the stem-loop(s) is shown in the expanded view below their cognate DGR cassette. As illustrated, stem-loops consist of a 6-8 bp stem and a 3-4 nt loop composed of the conserve sequence GNA or GRNA (where R= A or G, N= any nucleotide) (b) DGR stem-loops from disparate species support BPP-1 DGR mutagenic homing. The graph shows phage tropism switching frequency (TSF) observed for BPP-1 Δ ATR carrying a WT or stem-loop sequences from bacterial or archaeal DGRs shown in a. Donor RT indicates pMX-1 plasmid expressing a functional RT (+) or an enzymatically inactive RT (-). No phage plaques were detected for RT-deficient donors and the TSF shown represent limits of detection. (c) Size and nucleotide composition of the loop contributes to target recognition in BPP-1 DGR mutagenic homing. The graph shows TSF observed for BPP-1 Δ ATR carrying a WT, the *L. pneumophila* D5591 stem-loop (LPGCA) or a stem-loop sequence containing a single nucleotide loop insertion (LPGCXA, where X= A, G, C or T). Bars represent average TSF values from three independent experiments, and error bars represent \pm SEM. Asterisks indicate P values comparing BPP-1 carrying stem-loop sequences from disparate species to WT in Student's T tests. *P<0.05.

Figure 7. Sequence requirements for stem-loop function in target site recognition.

The function of DGR stem-loops in target recognition is proposed to be determined by 1) the size and nucleotide composition of the loop (5'-**GNNX**-3' or 5'-**GNX**-3' where N=A, C, T or G and X= A or C), 2) a G/C-rich stem with a C.G base pair (bold) adjacent to the loop and 3) orientation of the stem-loop/cruciform structure (relative to VR).

Supplemental figure 1. Sequence analysis of tropism switching products from

BPP-1 Δ ATR phage carrying WT stem-loops. Analysis of VR sequences from five (Wt1-5) progeny phages with switched tropisms indicates mutagenesis at VR positions (1-154) corresponding to adenines in the cognate TR. The region corresponding to the WT stem-loop is underline in black.

Supplemental figure 2. Sequence analysis of tropism switching products from

BPP-1 Δ ATR SLM2A phage. Mutagenesis at VR positions (1-54) corresponding to adenines in TR is observed in five (SLM2A1-5) progeny VRs from phage with switched tropisms. Sequence corresponding to the SLM2A stem-loop mutant is underline in black.

Supplemental figure 3. Sequence analysis of tropism switching products from

BPP-1 Δ ATR SLM2B phage. Mutagenesis at VR positions (1-54) corresponding to adenines in TR is observed in all five (SLM2B1-5) progeny VRs from phage with switched tropisms. Sequence corresponding to the SLM2B stem-loop mutant is underline in black.

Supplemental figure 4. Sequence analysis of tropism switching products for BPP-1ΔATR SLMGC phage. Mutagenesis of nucleotide positions in VR (1-54)

corresponding to adenines in TR is observed in all five (SLMGC1-5) progeny VRs from phage with switched tropisms. Sequence corresponding to the SLMGC stem-loop mutant is underline in black.

Supplemental figure 5. Sequence analysis of PCR-based homing products.

(a) Sequence alignment of VRInv homing products (VRInv1-5) with the predicted homing product of the nonmutagenized VR sequence (WTVR5'end). Sequences corresponding to the 5' end of VR to the end of the PCR tag are shown. VRInv1-5 are independent clones of progeny VRs from a single homing assay. The PCR tag is detected in all homing products, and adenine mutagenesis is observed in 2 of the 5 homing products. The primer P4 annealing site is indicated as a horizontal arrow. (b) Sequence alignment of VRInv homing products (VRInv1-5) with the predicted homing product of the nonmutagenized VR sequence (WTVR3'end). Sequences corresponding to the beginning of the PCR tag to the end of VR are shown. Adenine mutagenesis is observed in 3 of the 5 homing products. The primer P3 annealing site is shown as a horizontal arrow.

Figure 1.

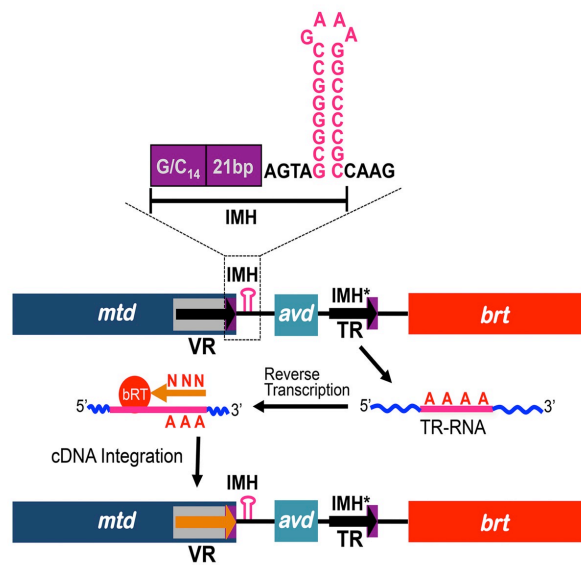
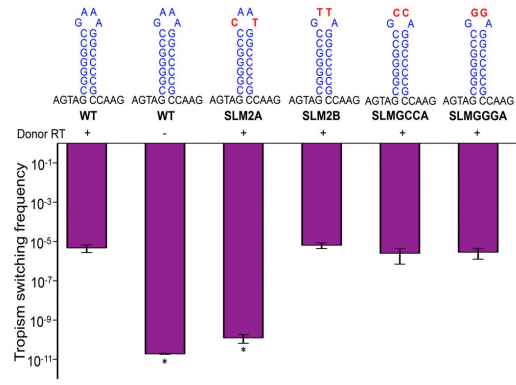


Figure 2.

A



B

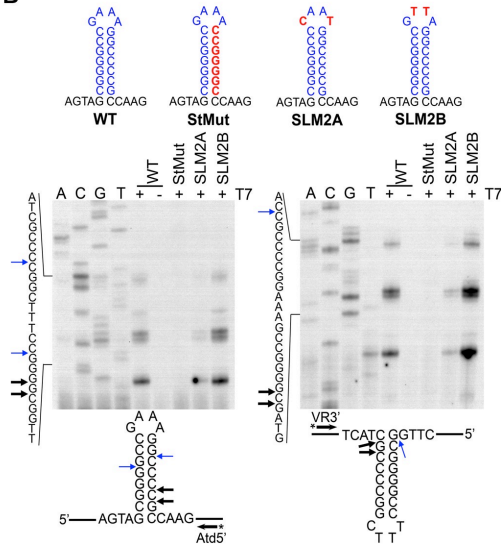
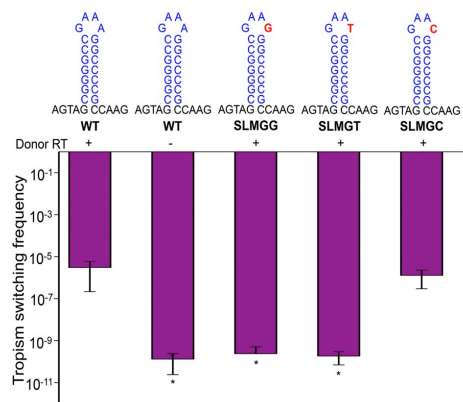


Figure 3.

A



B

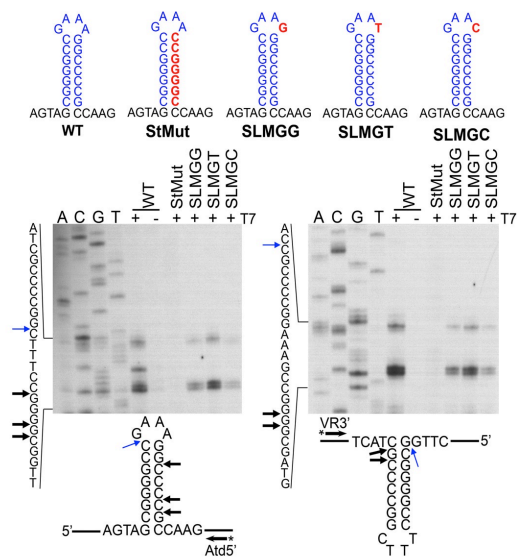


Figure 4.

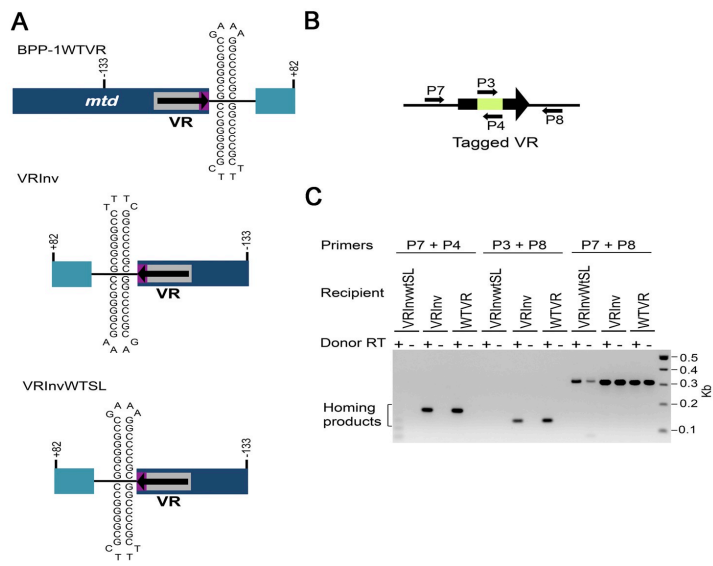


Figure 5.

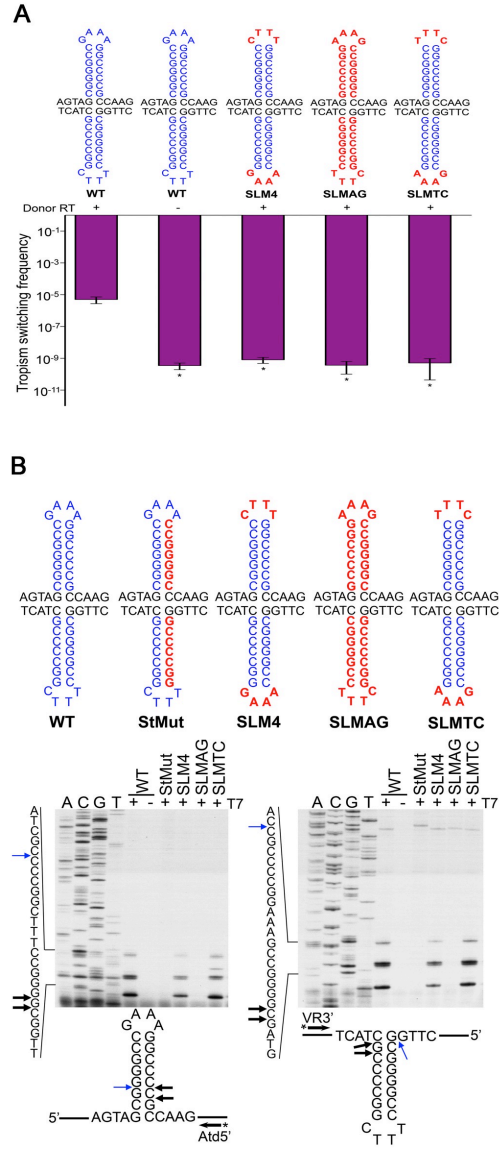


Figure 6.

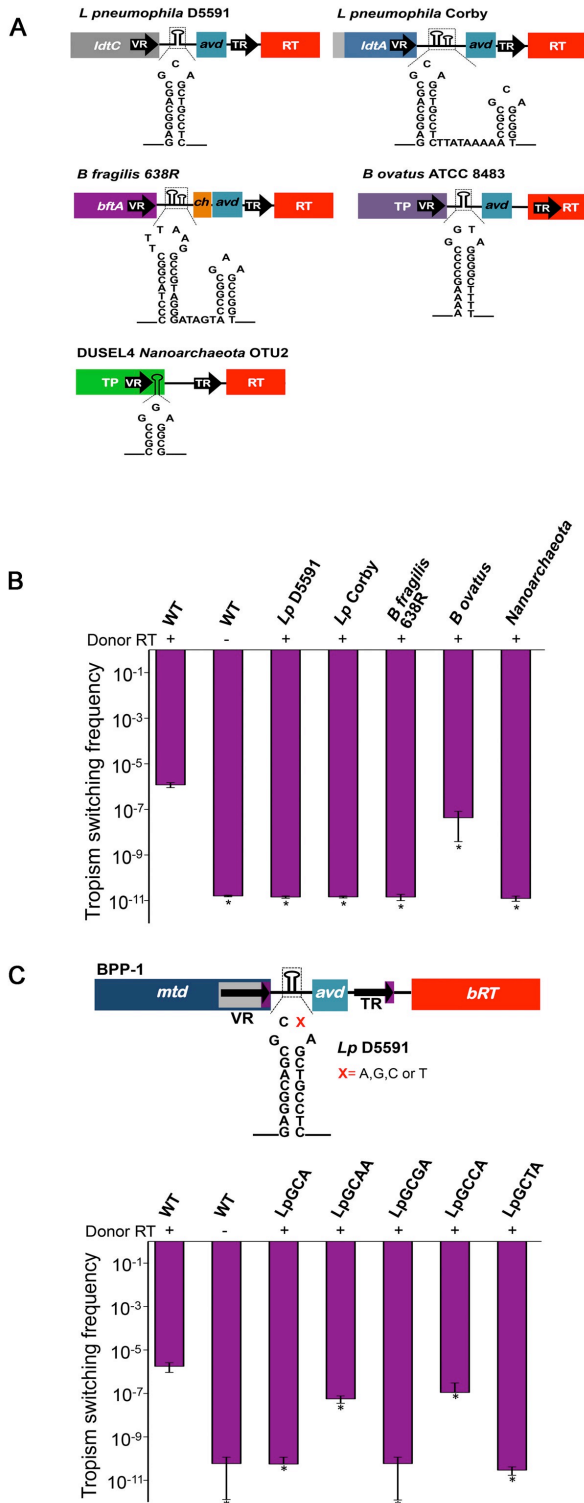
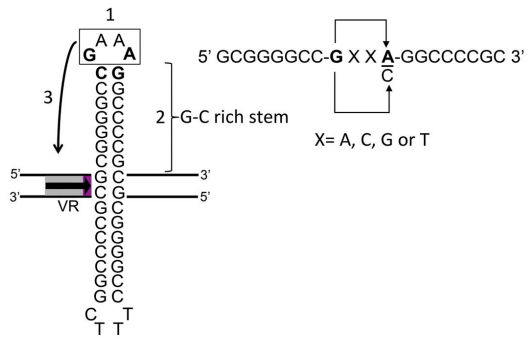


Figure 7.



Sup. Figure 1.

```

TR          1
CGCTGCTGCGCTATTCGGCGGCAACTGGAACAACAGTCGAACTCGGGTTCTCGCGCTGC
Wt1        CGCTGCTGCGCTATTCGGCGGCCTCCTGGAAACAACAGTCGTTCTCGGGTTCTCGCGCTGC
Wt2        CGCTGCTGCGCTATTCGGCGGCCTGGAGCAACAGTCGTACTCGGGTTCTCGCGCTGC
Wt3        CGCTGCTGCGCTATTCGGCGGCCTGGAGCAACAGTCGAACTCGGGTTCTCGCGCTGC
Wt4        CGCTGCTGCGCTATTCGGCGGCCTCCTGGGCCAACAGTCGAACTCGGGTTCTCGCGCTGC
Wt5        CGCTGCTGCGCTATTCGGCGGCCTGGAAACAACAGTCGTACTCGGGTTCTCGCGCTGC
*****.****.*****:*****

TR          GAACTGGAACAACGGGCCGTCGAACTCGAACCGGAACATCGGGCGCGCGCGCTGTGA
Wt1        GTACTGGTACAGCGGGCCGTCGTTCTCGTTCCGCTTCTTCGGGGCGCGCGCGCTGTGA
Wt2        GTACTGGAGCAGCGGGCCGTCGAACTCGTACGCGTCTTCGGGGCGCGCGCGCTGTGA
Wt3        GTACTGGTACTACGGGCCGTCGTACTCGTACGCGTACATCGGGCGCGCGCGCTGTGA
Wt4        GTACTGGAACAACGGGCCGTCGTACTCGCTCGCGTACCTCGGGCGCGCGCGCTGTGA
Wt5        GCTCTGGAACTACGGGCCGTCGCTCTCGAACGCGTCTTCGGGGCGCGCGCGCTGTGA
* :***:..:***** :*** :***:;* *****

          154
TR          CCACCTGATTCTTGAGTAGCGGGCCGAAAGGCCCGCCAAAGCAACCGATG
Wt1        CCACCTGATTCTTGAGTAGCGGGCCGAAAGGCCCGCCAAAGCAACCGATG
Wt2        CCACCTGATTCTTGAGTAGCGGGCCGAAAGGCCCGCCAAAGCAACCGATG
Wt3        CCACCTGATTCTTGAGTAGCGGGCCGAAAGGCCCGCCAAAGCAACCGATG
Wt4        CCACCTGATTCTTGAGTAGCGGGCCGAAAGGCCCGCCAAAGCAACCGATG
Wt5        CCACCTGATTCTTGAGTAGCGGGCCGAAAGGCCCGCCAAAGCAACCGATG
*****

```

WT Hairpin

Sup.Figure 2.

```

1
TR      CGCTGCTGCGCTATTTCGGCGGCAGCTGGAAACAACACGTGAACTCGGGTTCTCGCGCTGC
SLM2A1 CGCTGCTGCGCTATTTCGGCGGCAGCTGGAAACAACACGTGAACTCGGGTTCTCGCGCTGC
SLM2A2 CGCTGCTGCGCTATTTCGGCGGCAGCTGGAAACAACACGTGAACTCGGGTTCTCGCGCTGC
SLM2A3 CGCTGCTGCGCTATTTCGGCGGCAGCTGGAAACAACACGTGAACTCGGGTTCTCGCGCTGC
SLM2A4 CGCTGCTGCGCTATTTCGGCGGCAGCTGGAAACAACACGTGAACTCGGGTTCTCGCGCTGC
SLM2A5 CGCTGCTGCGCTATTTCGGCGGCAGCTGGAAACAACACGTGAACTCGGGTTCTCGCGCTGC
*****.*****.*****.*****

TR      GAACTGGAACAACGGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLM2A1 GAACTGGAACAACGGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLM2A2 GTACTGGAACAAGCGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLM2A3 GAACTGGAACAACGGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLM2A4 GAACTGGAACAACGGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLM2A5 GTACTGGAACAAGCGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
*:*****:.*:.******:*****.***:.* *****

154
TR      CCACCTGATTCCTTGAGTAGCGGGGCCAAAGGCCCGCCAGGCCAACCGATG
SLM2A1 CCACCTGATTCCTTGAGTAGCGGGGCCAAAGGCCCGCCAGGCCAACCGATG
SLM2A2 CCACCTGATTCCTTGAGTAGCGGGGCCAAAGGCCCGCCAGGCCAACCGATG
SLM2A3 CCACCTGATTCCTTGAGTAGCGGGGCCAAAGGCCCGCCAGGCCAACCGATG
SLM2A4 CCACCTGATTCCTTGAGTAGCGGGGCCAAAGGCCCGCCAGGCCAACCGATG
SLM2A5 CCACCTGATTCCTTGAGTAGCGGGGCCAAAGGCCCGCCAGGCCAACCGATG
*****:.*:.******:*****.***:.* *****

```

SLM2A Hairpin

Sup. Figure 3.

```

1
TR      CGCTGCTGCGCTATTCGGCGGC AACTGGAACAACAGTCGAACTCGGGTTCTCGCGCTGC
SLM2B1 CGCTGCTGCGCTATTCGGCGGC CCGCTGGCCAAACAGTCGTTCTCGGGTTCTCGCGCTGC
SLM2B2 CGCTGCTGCGCTATTCGGCGGC CTTGGAGCAACAGTCGAACTCGGGTTCTCGCGCTGC
SLM2B3 CGCTGCTGCGCTATTCGGCGGC CTTGGAGCAACAGTCGAACTCGGGTTCTCGCGCTGC
SLM2B4 CGCTGCTGCGCTATTCGGCGGC CTTGGAGCAACAGTCGTTCTCGGGTTCTCGCGCTGC
SLM2B5 CGCTGCTGCGCTATTCGGCGGC CTTGGAGCAACAGTCGTTCTCGGGTTCTCGCGCTGC
*****.***.*****:*****

TR      GAACTGGAACAACGGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLM2B1 GTA CTGGTACTACGGGCCGTCGTTCTCGTTTCGCGTACTTCGGGCGCGCGCGCTCTGTGA
SLM2B2 GTA CTGGAACTACGGGCCGTCGGCCCTCGTACGCGTTCA TCGGGGCGCGGGCGTCTGTGA
SLM2B3 GTA CTGGAGCAGCGGGCCGTCGTTCTCGTACGCGTACA TCGGGGCGCGGGCGTCTGTGA
SLM2B4 GTA CTGGAGCAACGGGCCGTCGAACTCGGCCGCGTTCTTCGGGCGCGGGCGTCTGTGA
SLM2B5 GTA CTGGAGCAGCGGGCCGTCGAACTCGTTTCGCGTTCA TCGGGGCGCGGGCGTCTGTGA
*:*****.:*.:***** ** ** ***:*****

154
TR      CCACCTGATTCTTGAGTAGCGGGGCCGAAAGGCCCGCC AAGGCAACCGATG
SLM2B1 CCACCTGATTCTTGAGTAGCGGGGCCGTTAGGCCCGCC AAGGCAACCGATG
SLM2B2 CCACCTGATTCTTGAGTAGCGGGGCCGTTAGGCCCGCC AAGGCAACCGATG
SLM2B3 CCACCTGATTCTTGAGTAGCGGGGCCGTTAGGCCCGCC AAGGCAACCGATG
SLM2B4 CCACCTGATTCTTGAGTAGCGGGGCCGTTAGGCCCGCC AAGGCAACCGATG
SLM2B5 CCACCTGATTCTTGAGTAGCGGGGCCGTTAGGCCCGCC AAGGCAACCGATG
*****:*****:*****

```

SLM2B Hairpin

Sup. Figure 4.

```

1
TR      CGCTGCTGCGCTATTCGGCGGGCAACTGGAACAACAGTCGAACTCGGGTTCTCGCGCTGC
SLMGC1  CGCTGCTGCGCTATTCGGCGGGCGCCTGGAGCAACAGTCGTTCTCGGGTTCTCGCGCTGC
SLMGC2  CGCTGCTGCGCTATTCGGCGGGCGCCTGGAGCAACAGTCGAACTCGGGTTCTCGCGCTGC
SLMGC3  CGCTGCTGCGCTATTCGGCGGGTCCCTGGAACAACAAGTCGTAAGTCGGGTTCTCGCGCTGC
SLMGC4  CGCTGCTGCGCTATTCGGCGGGCGCCTGGGCCAACAGTCGAACTCGGGTTCTCGCGCTGC
SLMGC5  CGCTGCTGCGCTATTCGGCGGGTCCCTGGAACAACAAGTCGAGCTCGGGTTCTCGCGCTGC
*****.****.*****:*****

TR      GAACTGGAACAACGGGCCGTCGAACTCGAACGCGAACATCGGGGCGCGGGCGTCTGTGA
SLMGC1  GGCCTGGAACCTCGGGCCGTCGTTCCTCGTTCGCGTTTCATCGGGGCGCGGGCGTCTGTGA
SLMGC2  GTAATGGTACAGCGGGCCGTCGAACTCGTACGCGTTTCATCGGGGCGCGGGCGTCTGTGA
SLMGC3  GTAATGGAGCAACGGGCCGTCGTACTCGTTCGCGTTCATCGGGGCGCGGGCGTCTGTGA
SLMGC4  GTAATGGACTACGGGCCGTCGTTCCTCGTTCGCGTACATCGGGGCGCGGGCGTCTGTGA
SLMGC5  GTAATGGAGCTACGGGCCGTCGTACTCGTACGCGTACATCGGGGCGCGGGCGTCTGTGA
* .**** .* *****:****:****:*.*****

154
TR      CCACCTGATTCCTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGCAACCGATG
SLMGC1  CCACCTGATTCCTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGCAACCGATG
SLMGC2  CCACCTGATTCCTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGCAACCGATG
SLMGC3  CCACCTGATTCCTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGCAACCGATG
SLMGC4  CCACCTGATTCCTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGCAACCGATG
SLMGC5  CCACCTGATTCCTGAGTAGCGGGGCCGAAAGGCCCGCCAAAGCAACCGATG
*****.*****.*****

```

SLMGC Hairpin

Sup. Figure 5.

A

```

WTVR5'end  CGCTGCTGCGCTATTCGGCGGCAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAAGCTT
VRInv1     CGCTGCTGCGCTATTCGGCGGCAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAAGCTT
VRInv2     CGCTGCTGCGCTATTCGGCGGCAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAAGCTT
VRInv3     CGCTGCTGCGCTATTCGGCGGCAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAAGCTT
VRInv4     CGCTGCTGCGCTATTCGGCGGCAGGTCGTCTGCGTTTGTGTTCCCTGTGCTAAGCTT
VRInv5     CGCTGCTGCGCTATTCGGCGGCAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAAGCTT
*****_*_******
    
```

P4

B

```

WTVR3'end  TCTAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAGCCATCGGGGCGCGCGCCTGTGAC
VRInv1     TCTAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAGCCATCGGGGCGCGCGCCTGTGAC
VRInv2     TCTAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAGCCATCGGGGCGCGCGCCTGTGAC
VRInv3     TCTAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAGCCATCGGGGCGCGCGCCTGTGAC
VRInv4     TCTAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAGCCATCGGGGCGCGCGCCTGTGAC
VRInv5     TCTAGATCTGTCTGCGTTTGTGTTCCCTGTGCTAGCCATCGGGGCGCGCGCCTGTGAC
*****_*_******
    
```

P3

G/C

MATERIALS AND METHODS

Bacterial strains, phage and plasmid constructs

Bordetella bronchiseptica strains RB53, RB54 and BPP-1 Δ ATR lysogens have been described [1, 5, 9]. Target region and stem-loop modifications were introduced into the BPP-1 Δ ATR lysogen through allelic exchange [2, 9].

Plasmid vectors pMX1 and pMX1SMAA were used for tropism switching assays and have previously been described [5, 9]. Plasmid pMX Δ TR23-96 and its RT-deficient derivative were used for homing assays and have been described [9].

Plasmid vector pUC18 carrying the WT BPP-1 target (pUC-WT) or stem-loop mutant derivatives (pUC-SLM2A, pUC-SLM2B, pUC-SLMGCCA, pUC-SLMGGGA, pUC-SLMGC, pUC-SLMGT, pUC-SLMGG, pUC-SLM4, pUC-SLMAG, pUC-SLMTC) used for *in vitro* analysis of stem-loop formation were constructed by cloning target sequences from position -6 upstream of VR to +82 downstream of VR.

Phage production for tropism switching and DGR homing assays

Phage production was carried out by mitomycin C induction from BPP-1 Δ ATR lysogens as previously described [5, 9]. Briefly, BPP-1 Δ ATR lysogens harboring appropriate donor plasmids were grown overnight at 37°C in Luria-Bertani (LB) media containing 25 mg/ml of chloramphenicol, 20 mg/ml of streptomycin and 10 mM nicotinic acid to modulate to the Bvg- phase and prevent expression of DGR components from the P_{pha} promoter. An amount of cells equal to 1 ml of culture (OD₆₀₀ = 1.0) was pelleted, rinsed, and resuspended in 2.5 ml Stainer Scholte (SS) medium containing 25 mg/ml chloramphenicol and 20 mg/ml streptomycin. Cells were grown for 3 h at 37°C to

modulate bacteria to the Bvg⁺ phase and induce expression of DGR components from the P_{fha} promoter. Phage production was induced with 0.2 mg/ml mitomycin C for 3 h at 37°C. Progeny phages were harvested by chloroform extraction [5, 9].

BPP-1 phage tropism switching and PCR-based DGR homing assays

Phage tropism switching assays were performed as previously described [5, 9]. DGR homing assays were performed as previously described with minor modifications [5]. Briefly, PCR was used to detect transfer of the 30 bp tag from donor plasmid TR to chromosomal target VR sequence in BPP-1 Δ ATR lysogens using primers pairs: P3-AAATCTAGATCTGTCTGCGTTTGTGTT, P4-AGCAAGCTTAGCACAGGAACACAAACG, P7-CCCTCTAGAGCTCCGGTTGCTTGTGGACG, and P8-AGCAAGCTTCCTCGATGGGTTCCAT. PCR products were cloned (TOPO-TA cloning; Invitrogen) and sequenced to verify transfer of the tag from TR to VR.

***In vitro* analysis of stem-loop formation**

Plasmids containing the WT BPP-1 DGR target or stem-loop mutant derivatives were isolated from *E. coli* DH5 α pir cells by a commercial kit (Qiagen). Stem-loop formation in supercoiled plasmid DNA was analyzed as in Guo *et al.* [9].

UV absorbance spectroscopy

UV absorbance melting profiles were obtained at 260 nm on a Hewlett- Packard HP8453 diode-array spectrophotometer equipped with a Peltier thermal controller.

Melting measurements of each oligonucleotide (3 μ M) were performed in a 10mM sodium phosphate buffer solution containing 10mM NaCl, pH 7.4. Prior to absorbance measurements, samples were incubated at 95°C for 5 min followed by slow cooling to room temperature. The absorbance was monitored with a temperature ramp of 1°C /min while the samples were heated from 20 to 98°C.

REFERENCES

1. 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.
2. 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**(7007): 476-481.
3. Medhekar B, Miller JF. 2007. Diversity-generating retroelements. *Curr Opin Microbiol* **10**: 388-395
4. Paul BG, Bagby S, Czornyj E, Arambula D, Handa S, Sczyrba A, Ghosh P, Miller JF, Valentine DL. 2015. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nature Commun* **6**: 6585.
5. Guo H, Tse L, 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**(6): 813-823.
6. McMahon SA, Miller J, 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**(10): 886-892.

7. 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**(6): 1196-1207.
8. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. 2010. Three-dimensional structure of tropism-sitching *Bordetella* bacteriophage. *Proc Natl Acad U S A* **107**(9): 4347-4352.
9. Guo H, Tse L, Nieh AW, Czornyj E, Williams S, Oukil S, Liu VB, Miller JF. 2011. Target Site Recognition by a Diversity-Generating Retroelement. *Plos Genet* **7**(12): 1-16.
10. Varani G. 1995. Exceptionally Stable Nucleic Acid Hairpin. *Annu Rev Biophys Biomol Struct* **24**: 379-404.
11. Dai X, Greizerstein MB, Nadas-Chinni K, Rothman-Denes LB. 1997. Supercoil-induced extrusion of a regulatory DNA hairpin. *Proc Natl Acad U S A* **94**: 2174-2179.
12. Chou SH, Zhu L, Reid BR. 1997. Sheared Purine-Purine Pairing in Biology. *J Mol Biol* **267**: 1055-1067.
13. Chou SH, Chin KH, Wang AH. 2003. Unusual DNA duplex and hairpin motifs. *Nuc Acids Res* **31**(10): 2461-2474.
14. Bevilacqua PC, Blose J. 2008. Structures, Kinetics, Thermodynamics, and Biological Functions of RNA Hairpins. *Ann Rev of Phys Chem* **59**: 79-103.
15. Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. *Microbiol Mol Biol Rev* **74**(4): 570-588.

16. Hirao I, Ishida M, Watanabe K, Miura K. 1990. Unique hairpin structures occurring at the replication origin of phage G4 DNA. *Biochim. Biophys. Acta* **1087**(2): 199-204.
17. Srinivasaraghavan K, Zacharias M. 2011. Role of the closing base pair for d(GCA) hairpin stability: free energy analysis and folding stimulations. *Nuc Acids Res* **39**(19): 8271-8280.
18. Kuznetsov SV, Ren CC, Woodson SA, Ansari A. 2008. Loop dependence of the stability and dynamics of nucleic acid hairpins. *Nuc Acids Res* **36**(4): 1098-1112.
19. Senior MM, Jones RA, Breslauer KJ. 1988. Influence of loop residues on the relative stabilities of DNA hairpin structures. *Proc Natl Acad U S A* **85**: 6242-6246.
20. Chou SH, Tseng YY, Chu BY. 1999. Stable formation of a Pyrimidine-rich Loop Hairpin in a Cruciform Promoter. *J Mol Biol* **292**: 309-320.
21. Hirao I, Nishimura Y, Tagawa Y, Watanabe K, Miura K. 1992. Extraordinarily stable mini-hairpins: electrophoretical and thermal properties of the various sequence variants of d(GCGAAAGC) and their effect on DNA sequencing. *Nuc Acids Res* **20**(15): 3891-3896.
22. Hadden JM, Déclais AC, Carr SB, Lilley DM, Phillips SE. 2007. The structural basis of Holliday junction resolutions by T7 endonuclease I. *Nature Letters* **449**: 621-624.

23. Chang CY, Stellwagen NC. 2011. Tandem GA Residues on Opposite Sides of the Loop in Molecular Beacon-like DNA Hairpins Compact the Loop and Increase Hairpin Stability. *Biochemistry* **50**: 9148-9157.
24. Reiling C, Khutsishvili I, Huang K, Marky LA. 2015. Loop Contributions to the Folding Thermodynamics of DNA Straight Hairpin Loops and Pseudoknots. *J Phys Chem* **119**: 1939-1946.
25. 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 U S A* **110**(20): 8212-8217.
26. Hirao I, Kawai G, Yoshizawa S, Nishimura Y, Ishido Y, Watanabe K, Miura K. 1994. Most compact hairpin-turn structure exerted by a short DNA fragment, d(GCGAAGC) in solution: an extraordinarily stable structure resistant to nucleases and heat. *Nuc Acids Res* **22**(4): 576-582.
27. Chin KH, Chou SH. 2003. Sheared-type $G_{anti} \square C_{syn}$ base-Pair: A Unique d(GXC) Loop Closure Motif. *J Mol Biol* **329**: 351-361.
28. Blommers MJJ, Walters JALI, Haasnoot CAG, Aelen JMA, Van der Marel GA, Van Boom JH, Hilbers CW. 1989. Effects of base sequence on the loop folding in DNA hairpins. *Biochemistry* **28** (18): 7491-7498.
29. Glucksmann-Kuis MA, Dai X, Markiewicz P, Rothman-Denes LB. 1996. *E. coli* SSB Activates N4 Virion RNA Polymerase Promoters by Stabilizing a DNA Hairpin Required for Promoter Recognition. *Cell* **84**(1): 147-154

CHAPTER 4. Identification of Host-encoded Factors that Participate in DGR-mediated Mutagenic Homing

ABSTRACT

Diversity-generating retroelements (DGRs) are a family of retroelements that introduce targeted variability within DNA-encoded protein sequences. DGRs are unique among retroelements due to their potential to accelerate the evolution of adaptive traits through iterative rounds of protein diversification by a unique error prone reverse transcriptase-mediated process called mutagenic homing. Since their discovery in the *Bordetella* bacteriophage BPP-1, DGRs have been identified in numerous bacterial genomes and within members of Archaea and their viruses. Although DGRs are wide spread in nature and protein diversification has been demonstrated in both, phage and bacterial systems, the precise mechanism of DGR mutagenic homing remains to be elucidated. Here, we describe a *Bordetella bronchiseptica* genetic screen to identify host factors that play a role in mutagenic homing in the *Bordetella* phage BPP-1 DGR. To identify host-encoded factors that directly or indirectly influence DGR homing, a random transposon-insertion library was created. Individual transposon mutants were screened for insertions that had a significant effect on BPP-1 DGR homing as measured by quantitative Km resistance assays. As an alternative approach to identify host factors that contribute to DGR mutagenic homing, we performed targeted mutagenesis of candidate genes involved in DNA- and RNA-processing activities. Mutants were screened for their ability to support DGR activity using phage tropism switching assays and we identified mutations in genes encoding DNA- and RNA-processing enzymes that decreased tropism switching. The identification of host genes that impact BPP-1 DGR activity demonstrates that mutagenic homing involves both DGR-encoded and host-encoded factors that participate in the diversification of target proteins.

INTRODUCTION

Diversity generating retroelements (DGRs) are a distinct family of retroelements capable of generating massive amounts of variability within defined DNA sequences [1, 2]. Sequence diversification occurs through a reverse transcriptase mediated process in which nucleotide substitutions are introduced at specific sites within target genes [2, 3]. The prototypic DGR was discovered in the temperate phage BPP-1, which infects *Bordetella* species. [1]. BPP-1 DGR activity generates nucleotide variability in the target gene, *mtd* (major tropism determinant), which specifies phage tropism for host-cell receptors [1, 4, 5]. Diversification of *mtd* allows BPP-1 to recognize distinct host-cell surface molecules as ligands for infection [1]. According to our model (Figure 1), BPP-1 tropism switching results from the introduction of nucleotide substitutions in a variable repeat (VR) located at the 3' end of *mtd*. Sites of nucleotide substitutions in VR correspond to adenine nucleotides in a homologous template repeat (TR), which encodes an RNA intermediate that is reverse transcribed by the DGR-RT. During reverse transcription, adenine residues are replaced with random nucleotides and the resulting cDNA displaces the parental VR in a process termed mutagenic homing [6]. Mutagenic homing is proposed to occur through a unique target-primed reverse transcription (TPRT) mechanism similar to that of group II introns in bacteria [6]. We postulate that homing initiates with either a single strand nick or a double strand break in the IMH (Initiation of Mutagenic Homing) site, a DNA region located at the 3' end of VR. The resulting 3' hydroxyl serves as a primer to reverse transcribe the TR-RNA intermediate [6], thus generating a mutagenized cDNA product. Although the proposed model is consistent with present data, the mechanism of cDNA integration and the

function of IMH are unknown. We have demonstrated that mutagenic homing requires specific nucleotide sequences and structural elements, including target site recognition sequences [6, 7]. In the BPP-1 DGR, these requirements are provided by the IMH element which is composed of a 14 bp G/C stretch (G/C_{14}) followed by a 21 bp segment and an inverted repeat that forms a stem-loop or a cruciform structure with an 8 bp stem and a 4 nt loop [7]. We recently demonstrated that in addition to base pairing interactions in the stem, the specific sequence and length of the 4 nt loop are critical for DGR function [7].

More recently, *in vivo* and *in vitro* analyses of the stem-loop structure indicate that conserved loop residues contribute to stem-loop formation and stability, and are required for DGR mutagenic homing (chapter 3). In addition, the polarity of the stem-loop is required for structure formation, and the orientation of the loop nucleotide sequence plays a critical role in target site recognition during mutagenic homing. Similar stem-loops have been identified in the majority of DGRs, and our recent analysis indicated that these conserved elements are functionally interchangeable between disparate species and fundamental to target site recognition. Thus, we propose the stem-loop/cruciform structure serves as a recognition element for DNA processing events that lead to cDNA synthesis and integration. However, bioinformatic and functional analysis of DGRs reveals the lack of sequences predicted to encode DNA- or RNA-processing enzymes, and we hypothesize that *trans*-acting host factors play a pivotal role in mutagenic homing. For example, host proteins could cleave the target DNA to generate a 3' hydroxyl that could serve as a primer for reverse transcription. In addition, DNA cleavage needs to be repaired and might require the action of the host

DNA repair machinery. Furthermore, some of these factors could affect TR-RNA processing, adenine mutagenesis, or the activity of DGR components such as RT, Awd or IMH elements.

To randomly identify host-encoded factors that directly or indirectly influence the BPP-1 DGR homing, a transposon-insertion library was created. Individual transposon mutants were screened for insertions that had a significant effect on DGR homing as measured by a quantitative Km resistance assay. As an alternative approach to identify host factors, we performed targeted mutagenesis of host candidate genes involved in DNA- and RNA-processing activities. Host mutants were screened for their ability to support mutagenic homing using phage tropism switching assays and we identified DNA- and RNA-processing enzymes that participate in mutagenic homing.

RESULTS

Genetic screen to identify host proteins that influence the BPP-1 DGR homing

To identify host-encoded factors that directly or indirectly influence BPP-1 DGR activity, a transposon insertion library was generated in the *Bordetella bronchiseptica* (*Bb*) BPP-1 Δ *ATR-Kan*^S strain using the mTn5 <GmR> transposon system (Figure 2) [8]. Tn5 insertion mutants were analyzed by measuring their ability to serve as targets in quantitative kanamycin homing assays, which takes advantage of the ability of synthetic TRs to diversify synthetic VRs *in trans* (Figure 3) [8]. A recipient VR (VR-*Kan*^S), chromosomally integrated, encodes a kanamycin resistance allele (*aph3'la*) with a 3' truncation rendering it nonfunctional [7]. The donor plasmid expresses *avd*, *bRT* and a synthetic TR (TR-Km2) encoding the 3' end of the *aph3'la* open reading frame [7]. DGR-mediated retrohoming from the donor TR to the recipient VR-*Kan*^S repairs the *aph3'la* gene conferring kanamycin resistance (*Kan*^R) [7]. Mutagenized *Bb* BPP-1 Δ *ATR-Kan*^S lysogens carrying the VR-*Kan*^S cassette were transformed with donor plasmids, and individual mutants were plated on agar slabs containing kanamycin to detect homing activity. The number of *Kan*^R colonies compared to non-mutagenized *Bb* BPP-1 Δ *ATR-Kan*^S lysogens transformed with donor plasmids carrying an active or an enzymatically inactive *bRT* provided a measure of DGR homing efficiency (Figure 2). Screening of 1500 insertion mutants identified 5 potential mutants with increased DGR homing levels and 71 potential mutants with insertions that resulted in decreased DGR homing efficiency. Transposon integration sites were identified by arbitrarily primed PCR (Figure 4) [9, 10]. DNA flanking the Tn5 transposon was sequenced and insertion sites were mapped to the *Bb* RB50 genome using BLAST searches. Of the 76

candidate genes identified, nine were sites of multiple Tn5 insertions. To eliminate false positive candidate mutants that resulted from Tn5 insertions within DGR components in the recipient prophage or mutations that affected the expression of the Kan^R phenotype, mutant strains were screened for DGR homing efficiency using a plasmid-based kanamycin homing assay (Figure 5). A plasmid carrying a recipient VR-Kan^S cassette was transformed into individual mutant strains and the targeting efficiency was measured as described above. Of the 62 candidate mutants identified in the transposon library screen, only nine were shown to reproduce the initial phenotype. Of these nine mutants, three had an increase in DGR homing levels and six demonstrated a decrease in DGR homing efficiency compared to the positive control (Figure 6). Candidate genes were classified according to their GO (gene ontology) annotation on the EcoCyc website (<http://ecocyc.org/>) (Table 1) and are predicted to be involved in global transcriptional regulation (BB2526, BB0438), RNA processing (BB2525), and metabolism (BB4104, BB1243). The rest of the candidate genes encode a hydrolase (BB4800), a hypothetical protein (BB2505), a virulence sensor protein (BB2995) and one was located upstream of the LysR family transcriptional regulator, BB0438.

Candidate host factors that participate in BPP-1 DGR mutagenic homing

To corroborate the effects of candidate genes (Table 1) in DGR mutagenic homing, genes were independently mutagenized and their contribution to mutagenic homing was analyzed using phage tropism switching assays that quantitatively assess the relative frequency of DGR activity because it requires adenine mutagenesis during retrohoming [1, 2, 6, 7]. Single gene deletions were introduced into *Bb* BPP-1 lysogens, and the

ability of BPP-1 phage to switch tropism was tested. As a negative control, we used BPP-1 lysogens where the bRT had been deleted, and thus could not switch tropisms. Contrary to the phenotype observed in the initial screen and plasmid-based kanamycin homing assays, deletion of eight of these candidate genes (*ΔBB1243*, *ΔBB2305*, *ΔBB2525*, *ΔBB2526*, *ΔBB0438*, *ΔBB4104* and *ΔIR*) had no effect on tropism switching (Figure 7). The effect of deleting *bvgS* (BB2995) on tropism switching has yet to be determined. Collectively, the above findings indicate that candidate genes identified in the transposon-library screen are not directly involved in DGR mutagenic homing, and the initial phenotype observed for these candidate mutants might be due to indirect effects of transposon insertions.

Role of RNA-processing host factors in BPP-1 DGR mutagenic homing

As an alternative approach to identify host-encoded factors involved in BPP-1 DGR mutagenic homing, we performed targeted mutagenesis of a set of host genes that participate in RNA-processing events. DGR mutagenic homing is proposed to occur through a TPRT mechanism similar to that of group II introns in bacteria [11, 12, 13, 14]. Host-encoded RNA- and DNA-processing enzymes are proposed to be required for retrohoming of *Lactococcus lactis* L1.LtrB group II intron [15, 16, 17]. Therefore, we investigated whether similar RNA-processing enzymes participate in mutagenic homing. Single gene deletions were introduced into *Bb* BPP-1 lysogens, and their ability to support BPP-1 DGR mutagenic homing was tested using tropism switching assays [6, 7]. First, we investigated whether RNase E (*rne*), an RNA-processing enzyme that regulates RNA-degradation and is known to impede L1.LtrB group II intron retrohoming

[15], is involved in mutagenic homing. As shown in Figure 8, deletion of *Bb rne* (Δrne) resulted in a BPP-1 tropism switching frequency comparable to wild type (WT). In addition, the contribution of *Bb* RNase H1 (*rnhA*) and RNase H2 (*rnhB*) to mutagenic homing was investigated. RNAase H1 and H2 are predicted to specifically hydrolyze the RNA strand of RNA-DNA hybrids [18] and could participate in the removal of TR-RNA from RNA-DNA hybrids generated during reverse transcription. Moreover, RNase HI has been implicated in the degradation of the intron RNA from the RNA-DNA hybrid intermediate during L1.LtrB group II intron retrohoming in *E. coli* [16, 17]. Contrary to our prediction, the deletion of *Bb rnhA* ($\Delta rnhA$) had no effect on tropism switching (Figure 8), while the deletion of *rnhB* resulted in nonviable mutant cells. Moreover, deletion of host genes encoding putative endoribonucleases BB4737 and BB479 had no effect on tropism switching. However, deletion of the *Bb rnc* (Δrnc) which is predicted to encode RNase III resulted in a $\sim 10^2$ fold reduction in tropism switching frequency (Figure 8). This indicates that RNase III contributes to mutagenic homing. RNase III enzymes have been reported to hydrolyze double-stranded (ds) RNA and to participate in the processing of rRNA and some mRNAs [18, 19]. Thus, RNase III could affect TR-RNA processing (see discussion).

Contribution of DNA-processing host factors to BPP-1 DGR mutagenic homing

To investigate whether DNA-processing enzymes participate in DGR mutagenic homing, single gene deletions of *recJ*, *rep*, and *recQ* were introduced into *Bb* BPP-1 lysogens and the effects of deleting these host genes on BPP-1 DGR mutagenic homing was determined using tropism switching assays. The *Bb recJ* encodes a single

stranded exonuclease, and *recQ* and *rep* both encode an ATP-dependent DNA helicase. The exonuclease activity of the host-encoded ReJ appears to be required for L1.LtrB group II intron retrohoming following TPRT in *E. coli* [15, 17]. As shown in Figure 9, mutations in *recJ* ($\Delta recJ$) and *rep* (Δrep) had no effect on tropism switching, indicating that they do not participate in mutagenic homing. However, the deletion of *recQ* ($\Delta recQ$) resulted in a $\sim 10^2$ fold reduction in tropism switching frequency which indicates that the host-encoded RecQ helicase contributes to mutagenic homing.

DISCUSSION

The BPP-1 DGR has been studied in mechanistic detail and a model for DGR-mediated mutagenic homing has been proposed based on present data. However, the precise mechanism of mutagenic homing remains inconclusive, and the contribution of host-encoded factors to DGR activity has not been previously analyzed.

Here, we used a genetic approach to identify host-encoded factors that participate in DGR mutagenic homing. First, we generated a Tn5 transposon library to identify host-encoded factors that directly or indirectly influence DGR activity. Although the initial transposon screen identified several candidate mutants, the function of the presumptive candidate genes in mutagenic homing could not be verified by subsequent analyses, indicating that candidate mutants were likely false positives. A subset of these presumptive hits showed a decreased in DGR activity which might be due to transposon insertions that affected the replication of the donor plasmid or the expression of the kanamycin resistant marker. Alternatively, the DGR activity initially observed in candidate mutants might be due to additional transposon insertions that were not accounted for. In some instances, transposon insertions were located in genes that are part of an operon and the effects of insertions on downstream gene expression remains to be addressed.

As an alternative approach to identify host-encoded factors that participate in BPP-1 DGR mutagenic homing, we performed targeted mutagenesis of a subset of host-encode RNA- and DNA-processing enzymes. Among five of the RNA-processing

enzymes analyzed, RNase III (encoded by *rnc*) was found to contribute to mutagenic homing. In *E. coli* RNase III hydrolyzes dsRNA and participates in the maturation of rRNA and some mRNAs [18, 19]. In addition, RNase III has been reported to mediate the degradation of a number of cellular and phage mRNAs [18, 19]. Given the function of RNase III in RNA metabolism, it is possible that RNase III is involved in TR-RNA processing or influences TR-RNA stability during mutagenic homing.

Furthermore, our analysis indicated that RecQ plays a role in DGR mutagenic homing. RecQ DNA helicases have been reported to participate in multiple cellular processes, including DNA recombination and repair [20, 21]. Moreover, RecQ DNA helicases facilitate the unwinding of double-stranded DNA and are implicated in the resolution of DNA structures [20, 21]. In *Neisseria gonorrhoeae*, RecQ binds and unwinds a quaduplex DNA structure required for efficient levels of pilin antigenic variation, a mutagenic system use by *N. gonorrhoeae* to evade the host immune response during infection [22]. We hypothesize that RecQ unwinds the DGR stem-loop/cruciform structure to facilitate the recognition of adjacent nucleotide sequences by host- or DGR-encoded factors that participate in cDNA integration.

The identification of host genes that impact BPP-1 DGR activity demonstrates that mutagenic homing involves both DGR-encoded and host-encoded factors, providing insight into a highly conserved mechanism through which all DGRs are predicted to function.

Figure legends:

Figure 1. *Bordetella* phage BPP-1 DGR mutagenic homing model.

In the current model, BPP-1 phage DGR mutagenic homing occurs through a target-primed reverse transcription mechanism (TPRT) [6]. DGRs diversify DNA sequences through a process called mutagenic homing, which introduces nucleotide substitutions into the VR (green arrow) of the target gene, *mtd* (green). Mutagenic homing initiates with either a single strand nick or a double-stranded break in the IMH (pink), a DNA region located at the 3' end of VR, and the resulting 3' hydroxyl serves as a primer to reverse-transcribe the TR-derived RNA intermediate (blue). The TR-RNA provides a template for DGR-RT (red box) dependent cDNA synthesis. During reverse transcription, TR-adenines (A, red) are randomly changed to any of the four nucleotides (N, red), which result in a mutagenized cDNA that displaces the parental VR. In the BPP-1 DGR, the target gene encodes the phage tail fiber protein responsible for binding to host-cell receptors (green circles) and mutagenic homing results in Mtd variants (red circles) that recognize new host-cell surface molecules as ligands for infection. Avd (aqua box) encodes an accessory protein that is proposed to interact with RT and nucleic acids [23].

Figure 2. Tn5 transposon library screen to identify host-encoded factors required for BPP-1 DGR activity.

Schematic of the DGR homing assay that was used to screen the Tn5 transposon library. Following conjugation, *Bb* BPP-1 Δ ATR*Kan*^S cells containing a randomly integrated <mTn5> transposon were recovered in selective medium (Sm^RGm^R) and

transformed *en masse* with the pMINI-Km2B (Cm^R) donor plasmid. Individual mutants were grown overnight in 24-deep well plates, and cells were pelleted, washed and resuspended in Stainer Scholte media to induce expression of DGR components from the P_{Fha} promoter. Cultures were induced for 6 hours and an equal number of cells were spotted onto 96-well agar slabs containing Kanamycin. The number of Kan^R colonies on each well was compared to those of non-mutagenized *Bb* BPP-1Δ*ATR*Kan^S lysogens transformed with a donor plasmid carrying an active bRT (+ control) or an enzymatically inactive bRT (- control). A sample slab is shown. The expanded view shows the number of colonies observed for positive and negative controls.

Figure 3. Donor and recipient constructs used for *Kan^R* gene targeting.

The donor plasmid (left), pMX-Km2, carries engineered TRs containing the last 36bp of the *Kan^R* ORF. The recipient cassette (Recipient Phage, right) is located between *attL* and *bbp1* of BPP-1Δ*ATR*. Engineered recipient VRs (VR-*Kan^S*) contain the *Kan^R* reporter gene with a deletion of the last 6 codons and are located upstream of the IMH (blue and green).

Figure 4. Identification of Tn5 transposon insertions in candidate host mutants.

Tn5 insertions in candidate mutants were mapped using an arbitrarily primed PCR. P1 and P3 are transposon-specific primers. P2 and P4 are arbitrary primers annealing to the *Bb* chromosome. Gels show PCR fragments generated in PCR1 (left) and PCR2 (right) from individual Tn5 insertion candidates. Individual bands indicate amplified DNA fragments containing transposon insertions.

Figure 5. Donor and recipient plasmids used for *Kan^R* gene targeting.

The donor plasmid (left), pMX-Km2, carries engineered TRs containing the last 36bp of the *Kan^R* ORF. The recipient cassette is located in the recipient plasmid (right).

Engineered recipient VRs (VR-*Kan^S*) contain the *Kan^R* reporter gene with a truncation of the last 6 codons and are located upstream of the IMH (blue and green).

Figure 6. Candidate transposon mutants with increased or decreased DGR activity.

The graph represent DGR-mediated *Kan^R* targeting efficiencies observed for candidate mutants using plasmid-based Kanamycin homing assays. Candidate mutant strains were transformed with recipient plasmids. Following induction for donor plasmid expression and *Kan^R* targeting, cells carrying donor and recipient plasmids were plated on selective (+Kan) or non-selective plates to determine the relative numbers of *Kan^R* colonies. The bars represent mean \pm standard deviations (s.d.). P values comparing mutants to a non-mutagenized strain in Student's t tests are indicated with asterisks.

* $P < 0.05$. No *Kan^R* colonies were detected for RT-deficient donors (Km2SMAA) and the numbers represent limits of detection.

Table 1. Candidate host genes that participate in BPP-1 DGR homing.

Bb candidate genes that had an effect on DGR homing efficiency are listed. Candidate genes were classified according to the GO (gene ontology) annotation. Candidate mutants with a decreased (down) or increased (up) *Kan^R* targeting efficiency compared

to non-mutagenized strains are indicated. DGR homing efficiencies observed in candidate transposon insertion mutants are shown.

Figure 7. Contribution of candidate genes to BPP-1 DGR mutagenic homing.

The graph illustrates the tropism switching frequencies (TSF) observed for BPP-1 phage obtained from ML6401 lysogens carrying gene deletions. The bars represent mean TSF \pm s.d. P values comparing mutants to WT in Student's t tests are indicated with asterisks. *P<0.05. No phage plaques were detected for RT-deficient donors and the TSF shown represents limits of detection.

Figure 8. RNA-processing enzymes that participate in BPP-1 DGR mutagenic homing.

The graph shows phage tropism switching frequencies observed for BPP-1 phage obtained from ML6401 lysogens carrying gene deletions ($\Delta BB4737$, $\Delta BB4793$, Δrne , Δrnc , $\Delta rnhA$). The bars represent mean TSF \pm s.d. P values comparing mutants to WT in Student's t tests are indicated with asterisks. *P<0.02. No phage plaques were detected for RT-deficient donors and the TSF shown represents limits of detection.

Figure 9. DNA-processing enzymes that participate in BPP-1 DGR mutagenic homing.

Tropism switching frequencies observed for BPP-1 phage obtained from ML6401 lysogens carrying gene deletions ($\Delta recJ$, $\Delta recQ$, Δrep). The bars represent mean TSF \pm s.d. Asterisks indicate P values comparing mutants to WT in Student's t tests. *P<0.02.

No phage plaques were detected for RT-deficient donors and the TSF shown represents limits of detection.

Figure 1.

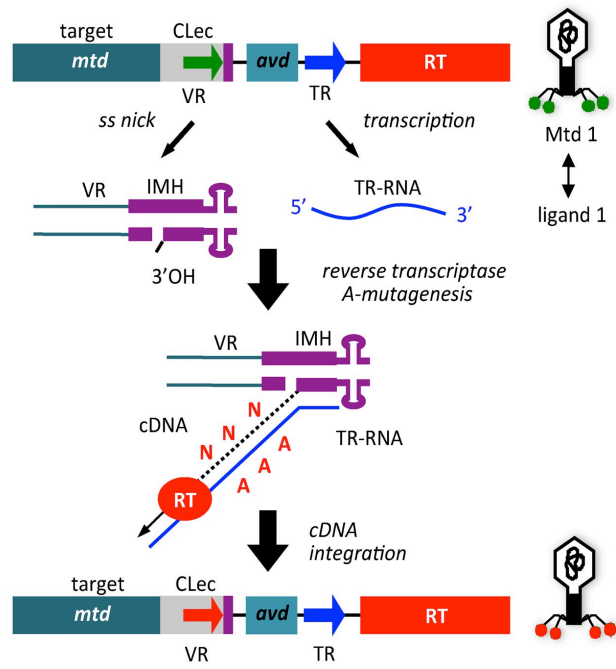


Figure 2.

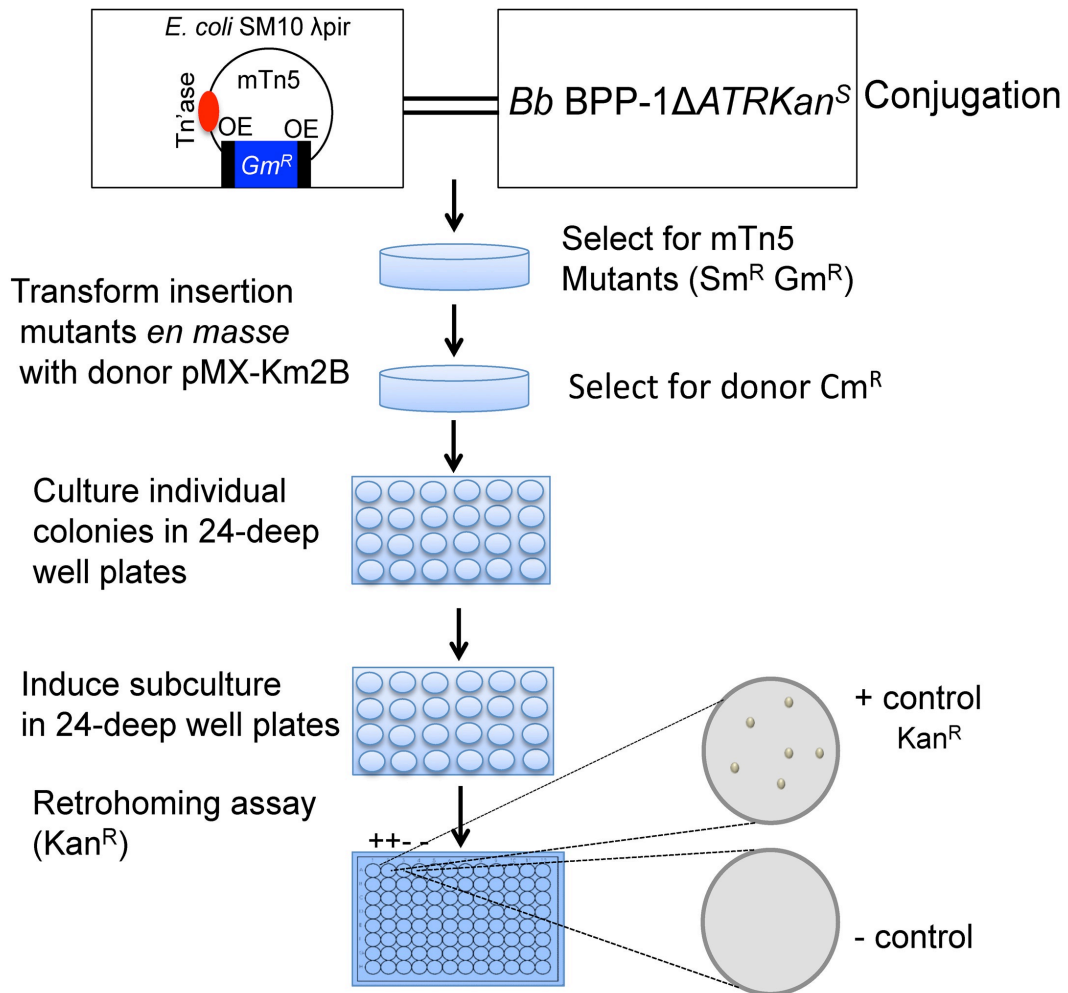


Figure 3.

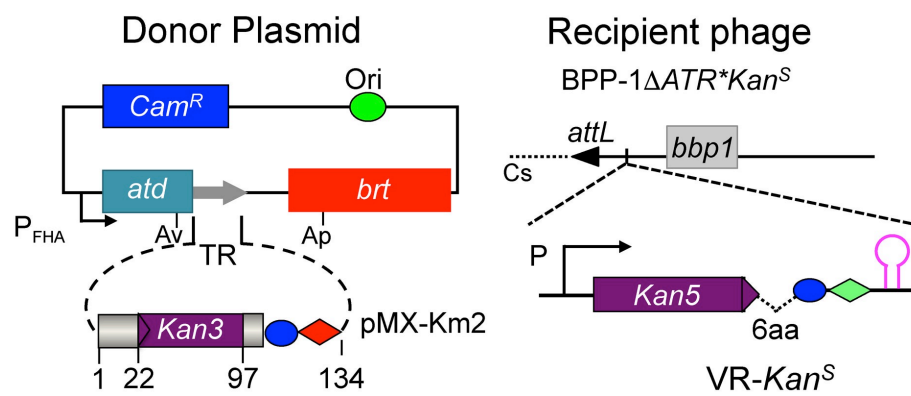


Figure 4.

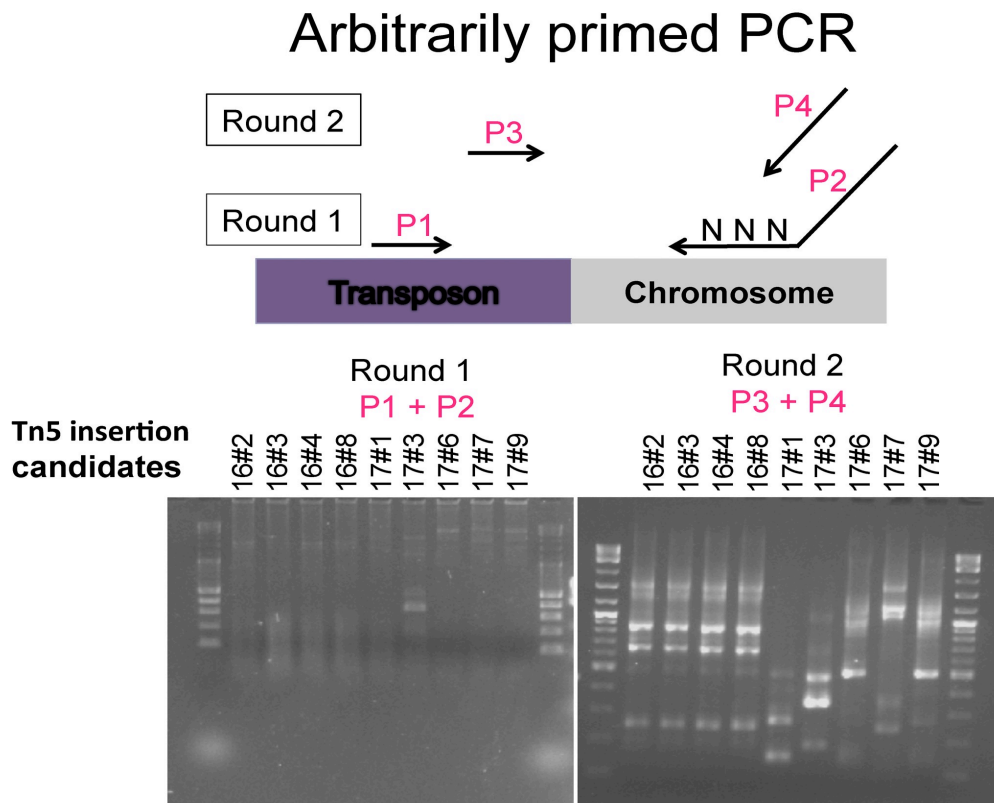


Figure 5.

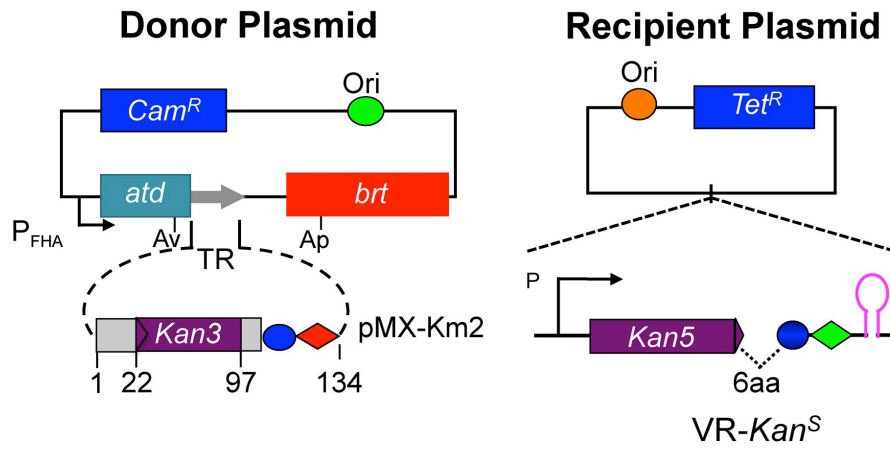


Figure 6

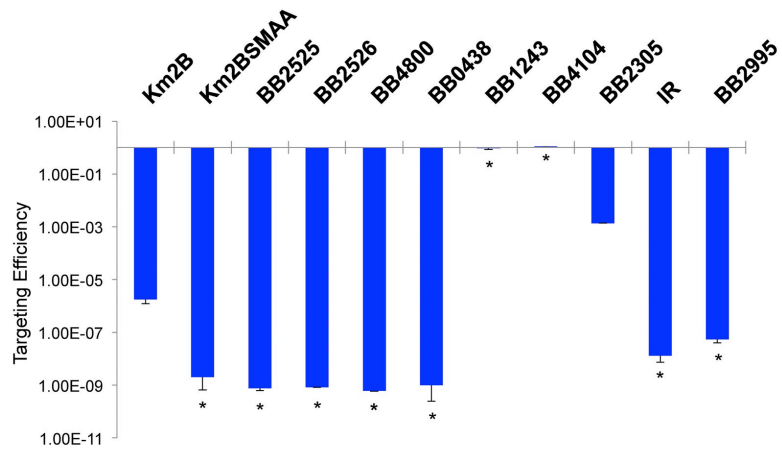


Table 1.

Classification	Gene disrupted	Downstream genes	Targeting efficiency Up/down	Kan^R Targeting efficiency
Global regulation	LysR family transcriptional regulator (BB2526)	None	Down	8.33E-10
	LysR family transcriptional regulator (BB0438)	None	Down	1.00E-09
RNA processing	RNA pseudouridylate synthase (BB2525)	None	Down	7.69E-10
Metabolism	Nucleotidyl transferase (BB4104)	None	Up	1.1E00
	ttrA; tetrathionate reductase subunit A (BB1243)	ttrB, ttrC	Up	9.90E-01
Unknown	Hydrolase (BB2305)	None	Up	1.37E-03
	Hypothetical protein (BB4800)	None	Down	6.25E-10
Other	BvgS; Virulence sensor protein (BB2995)	BvgR (BB2996)	Down	5.40E-08
	Intergenic region (between genes BB0437 and BB0438)	BB0438	Down	1.30E-08

Figure 7.

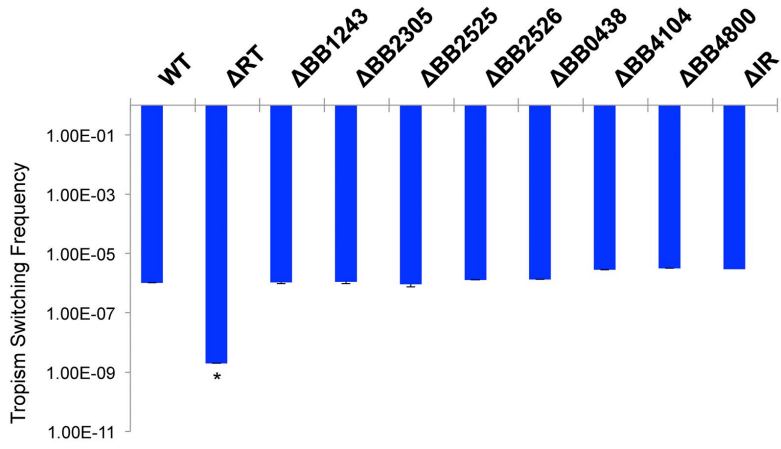


Figure 8.

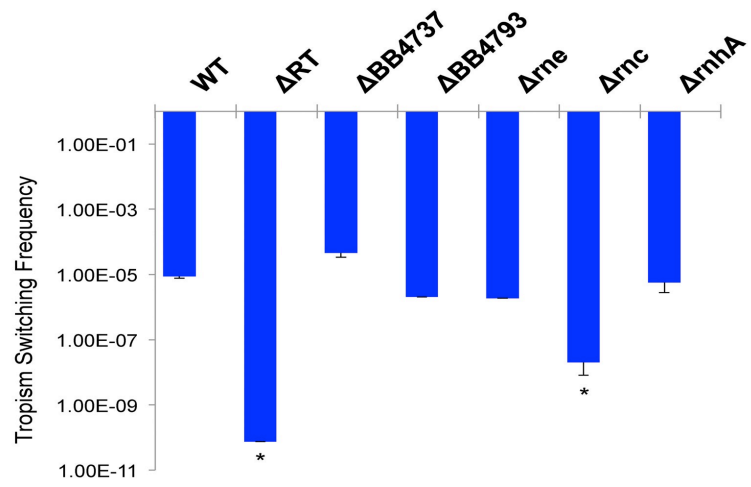
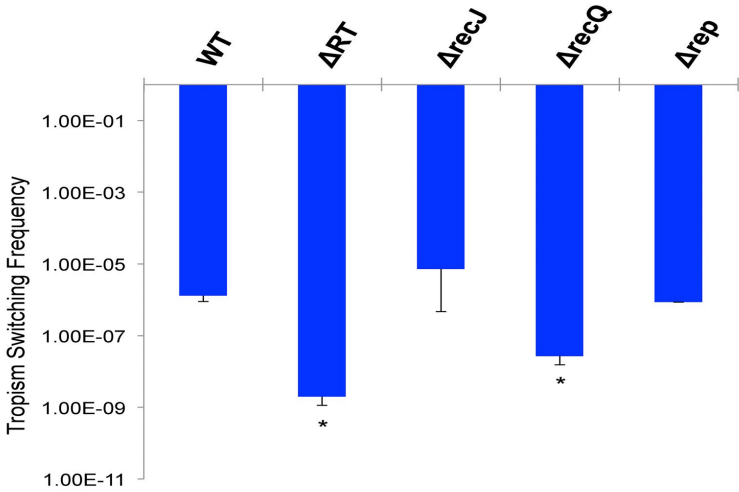


Figure 9.



MATERIALS AND METHODS

Bacterial strains

Bordetella bronchiseptica strains RB53 and RB54 have been previously described [1, 6, 7]. ML6401, a *Bordetella bronchiseptica* RB50 strain lysogenized with the BPP-1 phage has also been previously described [1, 6, 7]. The BPP-1 Δ ATR*Kan*^S lysogen was used to generate the mTn5 transposon library and has been described [8]. Gene deletions were introduced into ML6401 through allelic exchange [2, 6].

Plasmids

Plasmids pMX-Km2 and pMX-Km2SMAA have been previously described [8]. Plasmid pHGT-*Kan*^S was used as the recipient plasmid in the plasmid-based Km assays and has been previously described [7].

Transposon-library screen and Kanamycin homing assay

A random transposon mutagenesis library was created in the *Bb* BPP-1 Δ ATR-*Kan*^S strain using the mTn5 <Gm^R> transposon system [8]. To target the *Kan*^R resistance gene on the recipient phage, BPP-1 Δ ATR-*Kan*^S mutagenized lysogens were transformed *en masse* with donor plasmid pMX-Km2 and the cells were plated on Luria Bertani (LB) medium containing 20 mg/ml streptomycin (Str), 25 mg/ml of chloramphenicol (Cam), gentamycin (Gm) and 10 mM nicotinic acid (NA) to modulate cells to the Bvg⁻ to prevent transcription of DGR components from the P_{fha} promoter. After transformation of donor plasmid, individual colonies were picked and grown in 500 μ l of LB+NA+Str+Cam+Gm in 24-deep-well plates and incubated overnight at 37°C. An

amount of cells equal to 1ml of culture ($OD_{600}=1.0$) was pelleted, rinsed, and resuspended in 500 μ l of Stainer Scholte (SS) medium [24] containing 20 mg/ml streptomycin (Str), 25 mg/ml of chloramphenicol (Cam) and gentamycin (Gm). Cultures were grown at 37°C for 6 hours to modulate bacteria to the Bvg⁺ phase to induce transcription of DGR components from the P_{fna} promoter. After induction, an aliquot of 5 μ l was spotted onto LB+NA+Str and LB+NA+Str+Kan (50 ug/ml) 96-well agar slabs to determine *Kan^R* gene targeting efficiency. Growth, dilution and spotting parameters were optimized to allow sensitive detection of increased or decreased targeting efficiency. *Kan^R* targeting efficiency for each transposon-insertion mutant was determined by the number of Kan^R colony forming units (cfu) on LB+NA+Str+Cam+Kan relative to non-mutagenized BPP-1 Δ ATR-*Kan^S* lysogens transformed with donor plasmid carrying an active (pMX-Km2) or inactive bRT (pMX-Km2SMAA). Transposon insertion mutants with decreased or increased targeting efficiency were isolated and transposon insertion sites were identified by arbitrarily primed PCR [9, 10]. PCR products were sequenced and insertion sites were mapped to the *B. bronchiseptica* RB50 genome using BLAST searches.

Plasmid-based Kanamycin homing assays

To target the *Kan^R* resistance gene on a plasmid, individual transposon-insertion mutants identified in the initial screen were transformed with the recipient plasmid pHGT-*Kan^S*. Cells carrying donor and recipient plasmids were grown overnight at 37°C in LB+NA+Str+Cam+Tet (5 μ g/ml). An amount of cells equal to 1ml of culture ($OD_{600}=1.0$) was pelleted, rinsed, and resuspended in 2.5 ml SS+Cam+Str+Tet and

grown at 37°C for 6 hours. Serial dilutions were plated on LB+NA+Str+Cam+Tet and LB+NA+Str+Cam+Tet+Kan (50 ug/ml) to determine *Kan^R* gene targeting efficiencies. *Kan^R* targeting efficiency was determined as the ratio of colony forming units on LB+NA+Str+Cam+Kan to those in LB+NA+Str+Cam.

Phage production for tropism switching assays

Phage production was carried out by mitomycin C induction from ML6401 as previously described [6, 7]. Briefly, ML6401 lysogens carrying gene deletions were grown overnight at 37°C in Luria-Bertani (LB) media containing 20 mg/ml of streptomycin and 10 mM nicotinic acid. An amount of cells equal to 1 ml of culture ($OD_{600} = 1.0$) was pelleted, rinsed, and resuspended in 2.5 ml Stainer Scholte (SS) medium containing 20 mg/ml streptomycin. Cells were grown for 3 h at 37°C. Phage production was induced with 0.2 mg/ml mitomycin C for 3 h at 37°C. Progeny phages were harvested by chloroform extraction [6, 7].

BPP-1 tropism switching assays

Phage tropism switching assays were performed as previously described [6, 7].

REFERENCES

1. 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.
2. Doulatov S, Hodes A, Dai L, Mandan 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.
3. Medhekar B, Miller JF. 2007. Diversity-generating retroelements. *Current Opinion in Microbiology* **103**: 88-395.
4. 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.
5. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. 2010. Three-dimensional structure of tropism-switching *Bordetella* bacteriophage. *Proc Natl Acad U S A* **107**(9): 4347-4352.
6. Guo H, Tse LV, Barbalat R, Sivaamnuaiphorn S, Xu1 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.
7. Guo H, Tse LV, Nieh A, Czornyj E, Williams S, Oukil S, Lieu V, Martin D, Miller JF. 2011. Sequence Requirements for Target-Site Recognition by a Diversity-Generating Retroelement. *PLoS Genetics* **7**: 12.

8. De Lorenzo V, Herrero M, Jakubzik U, Timmis KN. (1990). Mini-Tn5 Transposon Derivatives for Insertion Mutagenesis, Promoter Probing and Chromosomal Insertion of Cloned DNA in Gram-Negative Eubacteria. *J Bacteriol.* **172**: 6568-6572.
9. Caetano-Annoles G. 1993. Amplifying DNA with arbitrary oligonucleotide primers. *PCR Methods Appl.* **3**: 85-92.
10. O'Toole GA, Pratt LA, Watnick PI, Newman DK, Weaver VB, Kolter R. 1999. Genetic approaches to study of biofilms. *Methods Enzymol* **310**: 91-109.
11. Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**:545-554.
12. 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.
13. Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. *Ann Rev Genet* **38**: 1-35.
14. Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol.* **3**(8): a003616.
15. Coros JC, Piazza CL, Chalamcharla VR, et al. 2008. A mutant screen reveals RNase E as a silencer of group II intron retromobility in *Escherichia coli*. *RNA* **14**: 2634-2644.
16. Smith, D., Zhong, J., Matsuura, M., Lambowitz, A. M., & Belfort, M. 2005. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes & Development* **19**(20), 2477–2487.
17. Yao J, Truong DM, Lambowitz AM. 2013. Genetic and Biochemical Assays Reveal a Key Role for Replication Restart Proteins in Group II Intron Retrohoming. *PLOS Genetics* **9**(4): e1003469.

18. Kushener SR. Messenger RNA Decay. EcoSal Plus 2013.
doi:10.1128/ecosalplus.4.6.4.
19. Zamore PS. 2001. RNA interference: listening to the sound of silence. *Nature Structural Biology* 8: 746-750.
20. Bernstein DA, Keck JL. 2003. Domain mapping of *Escherichia coli* defines the roles of conserved N- and C-terminal regions in the RecQ family. *Nuc Acids Res* **31**(11): 2778-2785.
21. Bernstein KA, Gangloff S, Rothstein R. 2011. The RecQ DNA helicases in DNA repair. *Annu Rev Genet.* 44: 393-417.
22. Cahoon LA, Manthei KA, Rotman E, Keck JL, Seifert HS. 2013. *Neisseria gonorrhoeae* RecQ Helicase HRDC Domains Are Essential for Efficient Binding and Unwinding of the pilE Guanine Quartet Structure Required for Pilin Antigenic Variation. *Journal of Bacteriology* **19**(10): 2251-2261.
23. 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**(2): 266-167.
24. Stainer DW, Scholte MJ. 1970. A simple chemically defined medium for the production of phase I *Bordetella pertussis*. *J Gen Microbiol* **63**: 211–220.

CHAPTER 5. Conclusion and Future Research

CONCLUSION

DGRs were discovered in the *Bordetella* phage BPP-1 and have since been identified in plasmids, bacteriophage and bacterial genomes, including members of the human microbiome, numerous pathogens of mammals or plants and organisms of environmental importance. More recently, DGRs were identified in Archaea and their viruses. Although all DGRs encode target proteins with diverse predicted functions, comparative bioinformatic analyses predict that all DGRs function by a conserved mechanism, and they represent a conserved prokaryotic system for targeted protein evolution.

Mechanistic studies of both bacterial and phage-encoded DGRs have provided insights into the mechanism by which diversification of target proteins occur. However, many questions remain unanswered regarding the precise mechanism for mutagenic homing. These include the exact function of the IMH element, how the stem-loop/cruciform structure formation contributes to target site recognition, or whether host-encoded factors participate in DGR mutagenic homing. The work presented in this thesis addressed some of these questions and provided insight into a conserved mechanism for DNA editing. As previously described in the preceding chapters, specific sequences and structural elements are required for target site recognition during mutagenic homing [1, 2, 3]. In the BPP-1 these requirements are provided by the IMH element, which is composed of a 14 bp G/C stretch (G/C_{14}), a 21 bp segment and an inverted repeat [1, 2, 3]. As shown in the second chapter of this thesis, the IMH inverted repeat forms a stem-loop or double-stranded cruciform structure with an 8bp stem and a 4nt loop. Formation

of the stem-loop/cruciform structure was confirmed by structure specific nuclease assays, and mutational analysis demonstrated that structure formation is required for efficient target site recognition. Characterization of the stem-loop/cruciform structure indicated that while the base pairing interactions in the stem are important for structure formation, the specific sequence and length of the 4 nt loop (GAAA) are crucial for DGR function. Furthermore, the analysis demonstrated that while the position of the stem-loop influences the efficiency of target site recognition, it does not affect cDNA integration at the 3' end of the target gene. This data indicated that formation and the location of the stem-loop/cruciform structure are necessary for efficient DNA sequence recognition of a target gene.

Furthermore, bioinformatics analysis revealed that stem-loops are common to the majority of DGRs. While the stem length varies between hosts, DGR stem-loops contain a GNA or GRNA loop (N=A, G, C or T and R=A, G). In Chapter 3 of this thesis, we investigated how loop nucleotide composition contributes to the structure formation and stability of DGR stem-loops. *In vitro* and *in vivo* analyses indicated that conserved G and A loop residues are essential to stem-loop/cruciform formation as well as thermal stability and consequently are required for mutagenic homing. In addition to influencing stem-loop/cruciform formation and stability, we demonstrated that orientation of the loop nucleotide sequence contributes to the directionality of mutagenic homing. This indicates that while they can serve as recognition elements on either DNA strand, the stem-loop/cruciform structures are functionally non-symmetrical and the directionality of mutagenic homing is likely determined by the loop nucleotide sequence. Stem-loop or cruciform structures are targets for many proteins [4, 5] and the polarity of RNA and

DNA stem-loops has been reported to allow the differential recognition of DNA strands by proteins that participate in replication, recombination and transcription regulation [4, 6, 7]. Thus, we hypothesize DGR stem-loop sequences are recognized either by the DGR AVD/RT complex or by host-encoded factors. For example, stem-loop/cruciform structures might be recognized by host DNA processing enzymes that participate in IMH cleavage, cDNA integration or DNA repair. Furthermore, due to the conservation of stem-loops in the majority of DGRs, including those recently identified in Archaea, we sought to analyze whether DGR stem-loops from bacterial and archaeal DGRs were able to support mutagenic homing in BPP-1. Here, we showed that DGR stem-loops are functionally interchangeable across species, which indicates that these conserved elements are fundamental to target site recognition. Together, these findings provide new insights that will help us uncover the precise function of the stem-loop/cruciform structure in target site recognition during mutagenic homing.

The notion that DGR stem-loops might serve as recognition elements for host-encoded factors that facilitate DNA processing events that culminate in cDNA synthesis, diversification and eventual integration, prompted us to investigate whether host-encoded factors participate in DGR mutagenic homing. In Chapter 4, we described a *Bordetella brochiseptica* (*Bb*) genetic screen to identify host factors that play a role in the BPP-1 DGR activity. To identify host factors that directly or indirectly influenced DGR activity, a transposon library was generated and over 1500 insertion mutants have been screened. Although many of these mutants initially appeared to either decreased or increased DGR activity, subsequent analysis indicated that candidate genes did not

participate in DGR homing. As an alternative approach to identify host factors, targeted mutagenesis of a set of host candidate genes that are known to participate in DNA- and RNA-processing events was performed. Our analysis revealed that the deletion of genes encoding RNase III and RecQ decreased tropism switching, which indicates that both DGR-encoded and host-encoded factors play a role in mutagenic homing. However, further analyses are needed to determine the action of these enzymes in mutagenic homing. In *E. coli*, the endoribonuclease RNase III is involved in RNA metabolism by specifically cleaving double-stranded RNA. Although RNase III primarily participates in the maturation of rRNA, it is also involved in the processing or decay of a subset of cellular and phage mRNAs [8, 9]. Given the role of RNase III in RNA metabolism, the *Bb* RNase III might be involved in TR-RNA processing, and this possibility will be analyzed in future studies.

Finally, our findings have implications for RecQ activity in mutagenic homing. RecQ ATP-dependent DNA helicases participate in DNA repair and have been reported to unwind DNA structures [10, 11, 12]. Accordingly, the DNA helicase activity of RecQ might facilitate the unwinding of the stem-loop/cruciform structure as local separation of DNA strands might allow the recognition of adjacent sequences by DGR conserved components or host-encoded proteins involved in cDNA integration at the 3' end of the target gene. To understand the role of RecQ in DGR mutagenic homing, we will test whether *Bb* RecQ interacts and unwinds the stem-loop/cruciform structure.

FUTURE RESEARCH

Analysis of the stem-loop function in target site recognition

We have demonstrated that stem-loop/cruciform structure formation is essential to DGR function and that target site recognition is determined by the loop nucleotide sequence (Chapter 2 and 3). Thus, we hypothesize that these conserved structures acts as recognition elements for conserved DGR components or host-encoded factors. To test this hypothesis, we will first determine whether the stem-loop/cruciform structure serves as a protein-binding element for Avd, RT or Avd-RT complex. His-tagged Avd and RT will be purified and combined with supercoiled plasmid DNA carrying wild type (WT) IMH sequences to determine if the stem-loop/cruciform element interacts with Avd, RT or Avd-RT *in vitro*. Complex formation will be determined by immunocoprecipitation assays using α Avd or α bRT antisera [13]. The presence of plasmid DNA in complexes will be measured by quantitative (q)PCR or by using radiolabeled DNA. The effects of stem-loop mutations on IMH recognition by transacting DGR components will also be analyzed. In addition, site-directed spin labeling (SDSL) combined with electron paramagnetic resonance (EPR) spectroscopy could be employed to rigorously monitor protein-stem-loop/cruciform interactions should they exist [14, 15]. Site-directed spin labeling has been used as a tool to analyze conformational changes that occur during transient tetraloop-receptor interactions [15]. Complex formation might be transient and therefore complex formation with WT IMH or IMH mutants with alteration of the stem-loop element will also be analyzed *in vivo* by chromosome immunoprecipitation [16].

As shown in Chapter 4, deletion of the *Bb recQ* decreased BPP-1 tropism switching which indicates the DNA helicase RecQ contributes to mutagenic homing. Since RecQ protein family members have been reported to bind and unwind DNA structures [10, 11, 12], we will test whether the *Bb* RecQ unwinds the DGR stem-loop/cruciform structure. First, we will analyze the interaction of RecQ with the stem-loop/cruciform structure *in vitro* by immunoprecipitation assays as previously described. Unwinding of the stem-loop/cruciform structure by RecQ will be determined by Circular Dichroism (CD) spectroscopy.

REFERENCES

1. 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**(7007): 476-481.
2. Guo H, Tse L, Barbalat R, Sivaamnuaiphorn 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**(6): 813-823.
3. Guo H, Tse L, Nieh AW, Czornyj E, Williams S, Oukil S, Liu VB, Miller JF. 2011. Target Site Recognition by a Diversity-Generating Retroelement. *Plos Genet* **7**(12): 1-16.
4. Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. *Microbiol Mol Biol Rev* **74**(4): 570-588.
5. Brazda V, Laister RC, Jagelska EB, Arrowsmith C. 2011. Cruciform structures are a common DNA feature for regulating biological processes. *BMC Mol Biol* **12**(33): DOI: 10.1186/1471-2199-12-33.
6. Dai X, Greizerstein MB, Nadas-Chinni K, Rothman-Denes LB. 1997. Supercoil-induced extrusion of a regulatory DNA hairpin. *Proc Natl Acad U S A* **94**: 2174-2179.
7. Glucksmann-Kuis MA, Dai X, Markiewicz P, Rothman-Denes LB. 1996.

- E. coli SSB Activates N4 Virion RNA Polymerase Promoters by Stabilizing a DNA Hairpin Required for Promoter Recognition. *Cell* **84**(1): 147-154.
8. Zamore PS. 2001. RNA interference: listening to the sound of silence. *Nature Structural Biology* **8**: 746-750.
 9. Kushener SR. Messenger RNA Decay. *EcoSal Plus* 2013.
doi:10.1128/ecosalplus.4.6.4.
 10. Bernstein DA, Keck JL. 2003. Domain mapping of Escherichia coli defines the roles of conserved N- and C-terminal regions in the RecQ family. *Nuc Acids Res* **31**(11): 2778-2785.
 11. Bernstein KA, Gangloff S, Rothstein R. 2011. The RecQ DNA helicases in DNA repair. *Annu Rev Genet.* **44**: 393-417.
 12. Cahoon LA, Manthei KA, Rotman E, Keck JL, Seifert HS. 2013. *Neisseria gonorrhoeae* RecQ Helicase HRDC Domains Are Essential for Efficient Binding and Unwinding of the pilE Guanine Quartet Structure Required for Pilin Antigenic Variation. *Journal of Bacteriology* **19**(10): 2251-2261.
 13. DA-Protein Interactions. 1996. *Current protocols in Molecular Biology*. 12.6.1-12.6.9.
 14. Hubbell WL, Cafiso DS, Atenbach C. 2000. Identifying conformational changes with site directed spin labeling. *Nature structural biology* **7**: 735-739.
 15. Qin PZ, Feigon J, Hubbell WS. 2005. Site-directed Spin Labeling Studies Reveal Solution Conformational Changes in A GAAA Tetraloop Receptor upon Mg²⁺-dependent Docking of a GAAA Tetraloop. *J. Mol. Biol.* **351**: 1-8.

16. Nelson JD, Denisenko O, Bomsztyk K. 2006. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nat Protoc* **1**: 179-85.
17. Yuann J-M P, Tseng W-H, Ling H-Y, Hou M-H. 2012. The effects of loop size on Sac7d-hairpin DNA interactions. *Biochimica et Biophysica Acta*. **1824**(9): 1009-1015.

APPENDIX A. Targeted diversity generation by intraterrestrial archaea and archaeal virus

ARTICLE

Received 10 Dec 2014 | Accepted 9 Feb 2015 | Published 23 Mar 2015

DOI: 10.1038/ncomms7585

OPEN

Targeted diversity generation by intraterrestrial archaea and archaeal viruses

Blair G. Paul¹, Sarah C. Bagby¹, Elizabeth Czornyj², Diego Arambula², Sumit Handa³, Alexander Sczyrba^{4,5}, Partho Ghosh³, Jeff F. Miller^{2,6,7} & David L. Valentine^{1,8}

In the evolutionary arms race between microbes, their parasites, and their neighbours, the capacity for rapid protein diversification is a potent weapon. Diversity-generating retroelements (DGRs) use mutagenic reverse transcription and retrohoming to generate myriad variants of a target gene. Originally discovered in pathogens, these retroelements have been identified in bacteria and their viruses, but never in archaea. Here we report the discovery of intact DGRs in two distinct intraterrestrial archaeal systems: a novel virus that appears to infect archaea in the marine subsurface, and, separately, two uncultivated nanoarchaea from the terrestrial subsurface. The viral DGR system targets putative tail fibre ligand-binding domains, potentially generating $>10^{18}$ protein variants. The two single-cell nanoarchaeal genomes each possess ≥ 4 distinct DGRs. Against an expected background of low genome-wide mutation rates, these results demonstrate a previously unsuspected potential for rapid, targeted sequence diversification in intraterrestrial archaea and their viruses.

¹Marine Science Institute, University of California, Santa Barbara, California 93106, USA. ²Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, California 90095, USA. ³Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California 92093, USA. ⁴Center for Biotechnology and Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany. ⁵DOE Joint Genome Institute, Walnut Creek, California 94598, USA. ⁶Molecular Biology Institute, University of California, Los Angeles, California 90095, USA. ⁷California NanoSystems Institute, University of California, Los Angeles, California 90095, USA. ⁸Department of Earth Science, University of California Santa Barbara, Santa Barbara, California 93106 USA. Correspondence and requests for materials should be addressed to D.L.V. (email: valentine@geol.ucsb.edu).

Energy-limited marine and terrestrial subsurface environments harbour a microbial reservoir of exceptional magnitude¹. Archaea are both numerically dominant² and well adapted to energy limitations faced in various intraterrestrial environments^{3,4}. Although little is understood about their physiology, metabolism, evolution, or mortality in these environments, current research predicts that they will be characterized by slow growth and low genome-wide mutation rates⁵.

Independent of the sporadic mutation rate, microbial genetic variation can be increased by processes such as gene conversion and horizontal gene transfer. The single most powerful such mechanism known in nature is the diversity-generating retroelement (DGR)^{6,7}. DGRs use a process called mutagenic retrohomology for the targeted replacement of a variable repeat (VR) coding region with a sequence derived from reverse transcription of a cognate non-coding template repeat (TR) RNA⁶⁻⁹. Crucially, the reverse transcriptase (RT) used is error-prone at template adenine bases¹⁰, but has high fidelity at other template bases, modulating the rate of diversification to permit rapid exploration of target protein (TP) variants within a recognizable structural framework. Over successive waves of replication, DGR activity leads to rapid evolution of TPs, typically altering ligand-binding specificity¹¹ and even permitting phage recognition of novel host ligands⁹. To date, DGRs have been found widely in bacteria and their viruses, but never in an archaeal system.

Because parasitism is expected to be an important driver of evolution and mortality in intraterrestrial archaea¹², we set out to identify and characterize viruses of anaerobic archaea from one system in the marine subsurface, a methane seep in a California borderlands basin. Our survey uncovers the complete genome of a virus that appears to infect archaea. Remarkably, this genome encodes a complete and apparently active DGR. We examine existing sequence data from archaeal systems, discovering multiple DGRs in the genomes of two subterranean nanoarchaea. These findings demonstrate that subsurface archaea and archaeal viruses maintain a mechanism for generating

protein hypervariability within targeted genes, bringing the capacity for massive diversification to the archaea-dominated deep biosphere.

Results

A putative archaeal virus encodes a DGR. We collected subsurface sediments from a methane seep at 820 m water depth in Santa Monica Basin. After confirming that these sediments exhibited anaerobic oxidation of methane (Supplementary Fig. 1), we prepared and sequenced a viral metagenome, uncovering a novel and apparently complete viral genome (termed ANMV-1; Fig. 1a). Examination of ANMV-1 coding sequences offered two key lines of evidence that this virus infects an archaeal host. First, the ANMV-1 genome encodes a TATA-box binding protein, an essential component of the transcriptional machinery in archaea and eukarya that is absent from bacteria¹³. Second, the ANMV-1 genome contains six genes that show sequence similarity (*e*-value 10^{-7} to 10^{-26}) with proteins from methanotrophic archaea (ANME-1 and ANME-2D) and none with comparable similarity to eukaryotic proteins (Supplementary Table 1). We further hypothesize that ANMV-1's archaeal host is anaerobic; ribonucleotide reductase activity is essential for phage genome replication¹⁴, and ANMV-1 encodes an oxygen-sensitive ribonucleotide reductase. In light of the active anaerobic oxidation of methane metabolism observed in the sample from which ANMV-1 was sequenced, the anaerobic archaeal host may belong to an anaerobic methane-oxidizing (ANME) clade.

Analysis of ANMV-1 identified a cassette bearing a RT gene, two 114-bp proximal repeats that vary from each other at positions corresponding to adenines, and a short inverted repeat with potential for hairpin formation (Fig. 1b). Together, these features are hallmarks of a DGR⁶⁻⁹. Since the discovery of these remarkable elements, >300 DGRs have been identified, all within the bacteria and their viruses^{15,16}. ANMV-1 represents the first identification of a DGR that appears to operate in an archaeal system.

Although the ANMV-1 VR lies within a gene of unknown function (best BLASTp *e*-value $>10^{-3}$, to uncharacterized proteins), the predicted secondary structure of the gene product

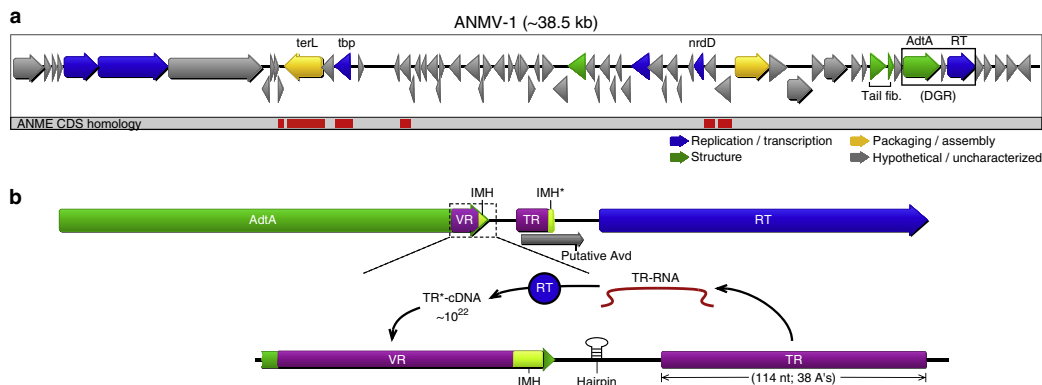


Figure 1 | Retroelement-containing ANMV-1 genome obtained from methane seep sediment. (a) Annotated coding sequences (CDS) designated by arrows that are coloured according to predicted function. Genes with blast similarity to ANME protein sequences are highlighted in red below each corresponding ANMV-1 locus (Supplementary Table 1). Symbols above selected annotations indicate putative gene names: terL, terminase large subunit; tbp, TATA-box binding protein; nrdD, anaerobic ribonucleoside triphosphate reductase; AdtA, DGR TP; RT, reverse transcriptase. An open box highlights the DGR cassette with flanking putative tail fibres (tail fib.), shown below the genome. (b) Putative *cis*- and *trans*-acting features of the ANMV-1 DGR. RT, accessory variability determinant (Avd) and AdtA ORFs are shown as blue, grey and green arrows, respectively. Purple boxes indicate template and variable repeat regions (TR and VR). The IMH and cognate IMH* sites are highlighted in yellow. The expanded DGR view depicts the putative retrohomology target site. Estimated number of nucleotide sequence variants is given above VR (TR* cDNAs), based on theoretical mutagenesis of adenines in TR intermediate RNA.

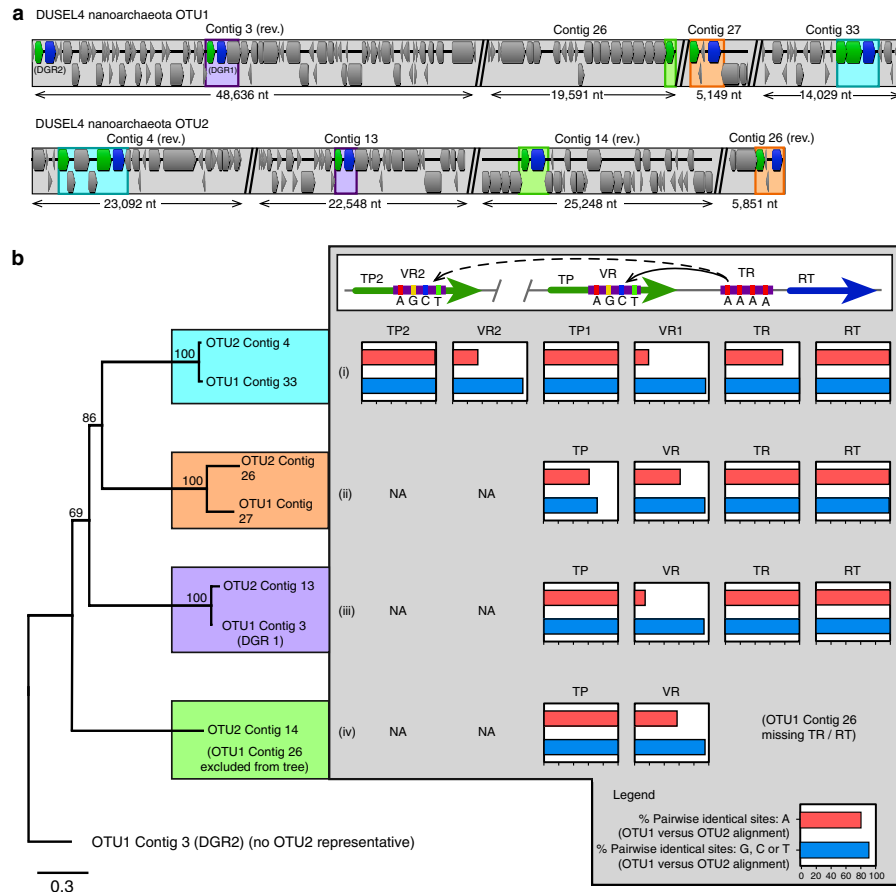


Figure 2 | Grouping of DGRs from Nanoarchaeota. (a) Positions of four DGR cassettes in each OTU, coloured by homology-based groups (note ungrouped OTU1 DGR in grey). Contigs are shown with DGRs on the forward strand (rev., reverse complement). (b) DGR groups, ordered by RT and TP homologies. A PhyML tree (left) was constructed with 100 bootstrap replicates (support indicated on branches) from concatenated alignments of TP and RT amino-acid sequences for each complete DGR cassette. Group 4 includes an incomplete DGR for OTU1 contig 26 (missing RT ORF). A schematic for nanoarchaeal DGRs shows the direction of information transfer during targeted mutagenesis. TP and RT genes are shown as green and blue arrows, respectively, while purple boxes indicate variable and template regions (VR and TR). Bar graphs show pairwise similarity between aligned OTU1 and OTU2 sequences for major DGR features, TP, VR, TR and RT. NA (not applicable) indicates that a feature is not found in the DGR.

offered important functional insights. The ANMV-1 DGR target (termed AdtA) shares greatest structural homology (37% of residues modelled with 99% Phyre confidence; r.m.s.d. 1.6 Å; $Z = 13.6$) with the major tropism determinant (Mtd) of *Bordetella* phage BPP-1, a DGR-targeted tail fibre protein responsible for binding host ligands. AdtA contains 21 codons with potential for adenine-specific amino-acid substitutions (versus 12 in Mtd), including nine AAY codons, with potential for $> 10^{18}$ variants. Thus, ANMV-1 demonstrates a degree of coding variability that is comparable to bacterial DGR systems¹¹ and outpaces the vertebrate immune system's capacity to generate variants of antibodies or T-cell receptor proteins^{17,18}. Predicted AdtA structural homology to Mtd is greatest in its C terminus, which corresponds to the C-type lectin (CLec)-fold common to many known bacterial DGR targets^{11,15}. As in Mtd, the targeted AdtA residues map to partially solvent-exposed sites in the CLec

domain (Supplementary Fig. 2). Together, these findings point to a binding-related role for AdtA, and the genomic proximity of the *adtA* gene to phage tail fibre genes (Fig. 1a) suggests host attachment as a possible function.

The discovery of a mechanism for rapid genetic diversification in ANMV-1 raises questions about the distribution and evolution of this virus. We conducted a search for close relatives of the ANMV-1 genome in environmental metagenomic databases, identifying a group of highly similar sequences (Supplementary Fig. 3) found in seafloor sediments of the Nyegga methane seeps, offshore Norway¹⁹, and in Coal Oil Point hydrocarbon seeps, offshore Santa Barbara, California. Metagenomes from both seeps cover portions of the ANMV-1 DGR cassette, including a closely related and intact RT open reading frame (ORF) from Nyegga seep sediments. These results indicate that ANMV-1 relatives are widespread in methane seeps. Furthermore, the persistence of

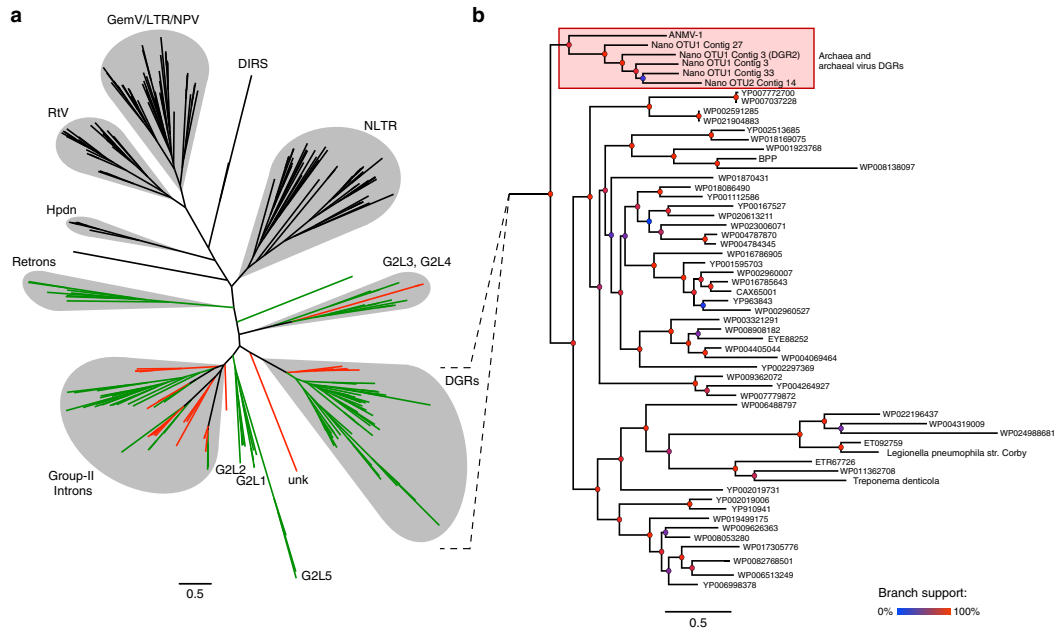


Figure 4 | RT phylogeny for archaeal DGRs. (a) Maximum-likelihood phylogenetic tree of RT representatives aligned with ANMV-1 and DUSEL4 Nanoarchaeota sequences. Green branches correspond to bacterial and bacteria-derived RTs (from chromosomes, plasmids, mitochondria, chloroplasts and bacteriophage), red branches indicate archaeal and archaeal virus RTs, and black branches represent RTs from eukaryotes and their viruses. Retroelement clades and key representatives are labelled as follows: DGRs, diversity-generating retroelements; DIRS, Dictyostelium retrotransposons; GemV, geminiviridae; G2L, group-II intron-like (G2L are numbered according to Simon and Zimmerly (24)); Hpdn, hepadnaviruses; LTR, long terminal repeat retroelements; NPV, nucleopolyhedroviruses; non-LTR, non-long terminal repeat retroelements; RIV, retroviridae; unk, unknown or unclassified. The scale shows substitutions per site. For clarity, bootstrap values are not shown for the full RT tree. **(b)** Expanded subtree view of DGR RT representatives. A red box highlights the archaeal DGR clade. NCBI accession codes are given for representatives in the subtree, but previously described bacterial DGRs are explicitly named. The representative for *Bordetella* phage BPP is labelled 'BPP'. Coloured circles at internal nodes indicate branch support.

across multiple DGR RTs in DUSEL4 (Supplementary Fig. 6a) suggests that they have a common source, perhaps a single acquisition followed by repeated gene duplications as new DGRs formed.

Nanoarchaeal DGRs target orphan genes. Most previously identified bacterial and phage DGRs diversify ligand-binding proteins, predominantly C-type lectin-like^{9,11,15} or immunoglobulin-like folds^{23,25}. By contrast, primary sequence analysis of all DUSEL4 *Nanoarchaeota* DGR and DGR fragment TPs reveals that they share no protein sequence homology with either AdtA or any database representatives, but rather constitute a set of orphan genes (Supplementary Fig. 6b); this finding is supported by Phyre analysis, which predicted no structural homology between characterized proteins and any nanoarchaeal TP. Initial structural investigation of one nanoarchaeal TP (OTU1 contig 3 DGR2 TP; Fig. 2b) by circular dichroism (CD) revealed that the purified protein adopts a thermostable fold ($T_m \sim 70^\circ\text{C}$; Supplementary Fig. 7) even with limited secondary structure (12% α -helix and 25% β -strand)²⁶. Pairwise sequence alignments of the nanoarchaeal TPs (Supplementary Fig. 6b) suggest that the targets of groups i–iv are unlikely to share substantial structural homology with each other, raising the possibility that nanoarchaeal DGRs may target a broader range of protein activities than are known for bacterial and phage DGRs.

Discussion

Comparison of the putative archaeal DGRs with the canonical bacterial and viral DGRs reveals both similarities and distinctive features that may influence DGR function. In *Bordetella* phage BPP-1, certain *cis*-acting elements appear critical for efficient retrohoming, including (1) an initiation of mutagenic homing (IMH) motif that lies at the 3' end of VR and an IMH* homologue at the 3' end of TR; and (2) a short inverted repeat downstream of VR, capable of forming a hairpin/cruciform structure, typically with a GRNA tetraloop¹⁰. DUSEL4 DGRs appear to maintain versions of these canonical *cis*-acting elements under additional constraints. First, IMH sites in DUSEL4 include a TGGGGT motif, while DUSEL4 IMH* sites carry a corresponding TGGAAT. Second, all DUSEL4 DGR hairpins have highly constrained GRA trinucleotide loops, and each hairpin lies within its DGR's TP gene, placing this region under selection at the level of both protein structure and DNA sequence. Investigation into the influence of these features on archaeal DGR activity may shed light on differences in the molecular mechanism of DGR retrohoming in bacterial and archaeal systems.

Examination of nanoarchaeal TRs suggests the capacity for individual DGRs to generate 7×10^{10} to 9×10^{12} variants of their TPs, with no risk of nonsense mutations (Supplementary Fig. 4). Although this range is low by comparison with typical bacterial and viral DGRs, the potential evolutionary impact must be

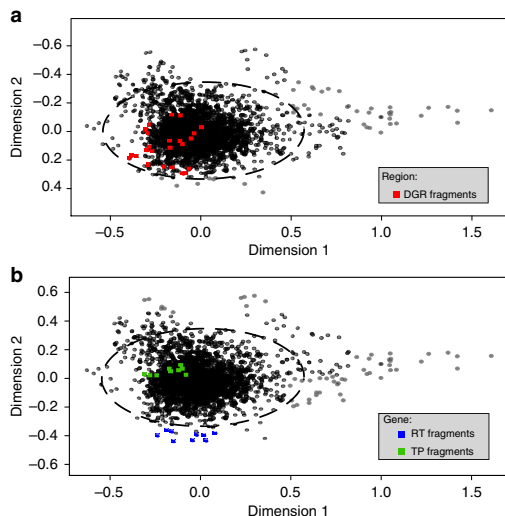


Figure 5 | Tetranucleotide distributions of DUSEL4 Nanoarchaeota. (a,b) Non-metric multidimensional scaling plots of tetranucleotide distributions of (a) concatenated DUSEL4 DGRs (red) and (b) separately concatenated DUSEL4 DGR RT (blue) and TP genes (green), compared with the rest of the DUSEL4 Nanoarchaeota OTU1 and OTU2 genomes (greyscale circles). Each point on the ordination plots represents one 5-kb fragment. Dashed ellipses indicate the 95% confidence region.

considered in light of the multiplicity of DGRs in DUSEL4 *Nanoarchaeota*; whereas no bacterial or viral genome has been found to harbour >2 distinct DGRs, these nanoarchaea have ≥ 4 . This profusion may enable subterranean nanoarchaea to explore a multidimensional fitness landscape far more rapidly than would sporadic mutation at the low rates observed for other intraterrestrial microbes⁵. Moreover, the fragmentary DGRs elsewhere in OTU1 suggest either that a single nanoarchaeal DGR can concurrently target multiple genes with homologous VRs, or that these DGRs are dynamic, with mobile RT/TR elements recruited from one locus to another over time. In either case, the diversity of nanoarchaeal DGR target sequences so far discovered raises the possibility that these organisms have used DGRs as a general tool for protein engineering—a hint that scientists might be able to do the same.

It is striking that these first discoveries of DGRs in archaeal systems should occur in a virus and in the *Nanoarchaeota*, a phylum associated with parasitism^{21,22}. Whether the uncultivated organisms represented by the DUSEL4 clade live as obligate parasites remains to be determined; their more important commonality with ANMV-1 may be their occurrence in Earth's subsurface. While massive and low-risk protein diversification offers clear advantages to any organism caught up in the Red Queen's race, the occurrence of a DGR in the globally distributed virus ANMV-1 and the proliferation of DGRs in subterranean nanoarchaea suggests that these elements may confer additional selective advantages in a compartmentalized and energy-limited subsurface environment.

Methods

Study site and sampling. Paull's Pingo is a seafloor mound feature (latitude 33.799° N and longitude 118.646° W, depth ~820 m) formed by the expansion of subsurface methane hydrate²⁷. We used active methane seeps at the pingo to collect sediment cores using deep submergence vehicle *Alvin*, during R/V *Atlantis*

Leg AT15-53 (September 2009). Sediment core processing was conducted shipboard in an anaerobic chamber, flushed with a nitrogen headspace. One sediment core was subsectioned between 5 and 15 cm (relative to seafloor) and dedicated to methane-amended incubations. Two subsamples of 60 ml sediment were homogenized with 20 ml of sterile, anoxic artificial seawater medium²⁸. Incubations with the homogenized sediments were prepared in 120-ml serum vials, under a 40-ml headspace of ~3% CH₄ and 97% N₂. Incubations were amended with ¹³C-labelled methane (99 atom-% ¹³C) as an exogenous tracer to track methane oxidation (Supplementary Fig. 1). Stable isotope ratios ($\delta^{13}\text{C}$) for CO₂ were measured by isotope ratio mass spectrometry (Thermo Finnigan Delta XP Plus in continuous flow mode). After 1 month of enrichment, the incubation was terminated and viruses were purified for DNA sequencing.

Virome purification and DNA sequencing. Incubation slurry samples (1:2 sediment:aqueous phase) were used for virus particle purifications. Samples were vigorously homogenized by vortexing (15 min), followed by centrifugation (10 min, 500g). Supernatant was filtered (0.22 μm) to separate viruses from cells. Viruses were concentrated and viral DNA was extracted as previously described²⁹. Briefly, virus particles were concentrated via caesium chloride density gradient ultracentrifugation (2 h, 22,000 g, 4 °C) and treated with DNase-I. DNA was extracted by cetrimonium bromide (CTAB)-chloroform and phenol-chloroform separation. Before viral DNA amplification, a 16S PCR assay to screen for cellular DNA contamination was performed with universal bacterial primers Bact27F (5'-AGAGTTTGATCCTGGCTCAG-3') and Bact1492R (5'-GGTTACCTTGTACGACTT-3'). Following this check, we performed Phi29 polymerase multiple displacement amplification (MDA) using the Illustra Genomiphi HY DNA Amplification Kit (GE Healthcare). Thermal cycling steps for denaturing template DNA, polymerase amplification, and post-amplification enzyme inactivation were performed according to the manufacturer's specifications, except that the MDA amplification reaction was incubated for 2 h instead of 4 h (2 h, 30 °C). Amplified product was pyrosequenced on 454-titanium plates at the Broad Institute, as part of the Moore Marine Phage Metagenome Initiative³⁰. Metagenomic reads can be obtained under the NCBI BioSample accession code PRJNA47435.DV-ANM1.

Read preprocessing, binning, and assembly. Raw sequencing reads were first scanned for sequencing primers, which were identified and removed using TagCleaner³¹. The reads were then preprocessed to remove low-quality sequence following the method of Hurwitz *et al.*³², using a custom R script. Preprocessing included, first, removal of any reads with ambiguous (N) bases; second, removal of the shortest 2.5% and longest 2.5% of reads; third, removal of reads with mean quality score >2 s.d. below the mean; and finally, de-replication with CD-Hit 454 (ref. 33).

Reads that passed preprocessing and quality control (QC) steps were subjected to *de novo* assembly using CAMERA's meta-assembler³⁴. As this assembler does not permit user manipulation of read overlap parameters, we compared the meta-assembler output with a custom reassembly approach using Geneious v7.0 (Biomatters Ltd) with the following parameters: minimum overlap 35 bases, overlap pairwise identity 90% and index word length 12 nt. The ANMV-1 contig described in this study was generated from the meta-assembly and aligned globally with 97.7% pairwise nucleotide similarity to a contig obtained by the second custom *de novo* assembly. PCR screening confirmed the authenticity of the ANMV-1 DGR cassette in both template and MDA-amplified viral DNA, using primers that partially overlap TP, RT and VR/TR regions: ANMVdgrF (5'-AGGCGATGCAGACGAATGGC-3') and ANMVdgrR (5'-TTGCCAGAGTTACCGGCG-3').

Metagenome annotations. Prediction of open reading frames was performed using Glimmer3 (ref. 35) with default parameters. Translated ORF sequences were annotated via CAMERA-HMM and BLASTp³⁶ searches against the following databases: TIGRfam, Pfam, COG and NCBI-nr (*e*-value < 10⁻³). To determine which ORFs from ANMV-1 genome share similarity to viral and prophage sequences, we compared our contig's translated ORFs with the ACLAME prophage-specific database³⁷. To assess similarity to proteins from anaerobic methane-oxidizing archaea, we inspected NCBI-nr BLASTp results for ANME protein hits (uncultured archaeon, ANME-1; *Candidatus* Methanoperedens nitroreducens', ANME-2D; and uncultured archaeon, Gfoz37D1). A BLASTn survey was conducted against environmental metagenomic databases, including NCBI metagenomic sequences (env_nt), Moore Marine Virus Metagenomes³⁰ and Pacific Ocean Virome sequences³⁸, to find representatives sharing high nucleotide similarity (*e*-value < 10⁻²⁰; 28-nt word size) with ANMV-1.

The putative DGR TP of ANMV-1, Adta, was analysed using Phyre2 (ref. 40) to find functional representatives based on secondary structural homology. Residues of TP that aligned with high confidence to the CLec fold region of the Mtd protein *Bordetella* phage BPP-1 (Phyre confidence >90%) were used to predict a three-dimensional model. Residue positioning was assessed by Ramachandran analysis and C-terminal variable residues were mapped from the primary sequence onto the predicted structure using Geneious v7.0 (Biomatters Ltd).

Comparative analysis of Nanoarchaeota genomes. We identified DGR-like RTs via BLASTp searches against the NCBI-WGS database. For an initial proxy of DGR repeat features, we used the EMBOSS tool Dotmatcher⁴⁰ to perform a dotplot analysis of homologous regions with moderate proximity (± 5 kb) to RT. TR/VR regions were confirmed from candidates that comprised mostly adenine-specific variability, with at least 10 adenine-specific mismatches, with respect to one strand, and no more than 2 non-adenine mismatches in 100 bp of aligned sequence.

DGR-containing sequences that were analysed in this study are from single-cell genomes belonging to DUSEL4 *Nanoarchaeota*, which were broadly described as part of a genome and metagenome annotation study on 'microbial dark matter', published elsewhere²⁰. DUSEL4 *Nanoarchaeota* representatives were previously assigned into two OTUs comprising four single-cell genomes. We describe *Nanoarchaeota* DGRs with reference to their occurrence in combined single-cell sequence assemblies: OTU1 (genomes AAA011-G17 and AAA011-L22) and OTU2 (genomes AAA011-J02 and AAA011-K22). To confirm the presence of multiple distinct DGRs in one single-cell genome, we aligned OTU1 sequences with contigs from *Nanoarchaeota* AAA011-G17, which has the highest genome completeness of the DUSEL4 representatives²⁰.

Nanoarchaeota RT sequences were aligned using ClustalW⁴¹ with sequences containing the catalytic RT domain, representing DGRs, group-II introns, retrons, long terminal repeats (LTRs), retroviruses, non-LTR elements and retroplasmids. The alignment was compared with a position-specific scoring matrix for the RVT-1 protein family (PF00078), and was manually realigned to conserve motifs considered essential for RT activity. Trees were constructed in MEGA v5.2 (ref. 42) using PhyML⁴² with the model LG + G + F. In addition, a PhyML tree was constructed from concatenated alignments of RT and TP amino-acid sequences to compare sequence similarities amongst Nanoarchaeota DGR cassettes.

TP expression and purification. Coding sequences of nanoarchaeal TPs were synthesized with codons optimal for expression in *Escherichia coli* (GENEWIZ, Inc.) and cloned into a modified pET28b expression vector with an N-terminal His-tag followed by a PreScission protease cleavage site. Construct integrity was confirmed by DNA sequencing. TPs were expressed in *Escherichia coli* BL21-Gold (DE3) cells. Bacteria were grown with shaking at 37 °C to an optical density (OD600) of 0.6–0.8 and then cooled to room temperature, followed by induction with 0.5 mM isopropyl β -D-1-thiogalactopyranoside. Bacteria were grown with shaking at room temperature for 5–6 h further, then harvested by centrifugation (25 min, 4,000g, 4 °C); the bacterial pellet was frozen at -80 °C.

Cells were thawed and resuspended in buffer A (300 mM NaCl, 50 mM Tris (pH 8) and 5 mM β -mercaptoethanol); 20 ml⁻¹ of bacterial culture) supplemented with 1 mM phenylmethylsulfonyl fluoride (PMSF). The bacteria were lysed by sonication and the lysate was centrifuged (30 min, 35,000g, 4 °C). The following steps were performed at 4 °C. The supernatant was applied to a column containing His-Select Nickel affinity gel (Sigma, 1 ml of resin per 20 ml of bacterial lysate), which had been equilibrated with buffer A. The column was washed with five column volumes of buffer B (300 mM NaCl, 20 mM Tris (pH 8) and 5 mM β -mercaptoethanol) containing 20 mM imidazole, and the TP was eluted with buffer B containing 250 mM imidazole. The His-tag was removed by PreScission protease cleavage (1:50 TP: protease mass ratio) overnight at 4 °C. Cleaved TP was separated from non-cleaved proteins by applying the sample to a His-Select Nickel affinity gel column (Sigma) and collecting the flowthrough. The TP was further purified by gel filtration chromatography (Superdex 75) in 300 mM NaCl, 20 mM Tris (pH 8) and 1 mM dithiothreitol. Purified protein was concentrated to 2 mg ml⁻¹ using ultrafiltration (10 kDa MWCO Amicon, Millipore); the concentration of TP was determined using a calculated molar extinction coefficient at 280 nm of 28,880 M⁻¹cm⁻¹.

CD spectroscopy. CD spectra were collected for the purified nanoarchaeal TP at 10 μ M in 300 mM NaF, 20 mM sodium phosphate buffer, pH 8, 1 mM dithiothreitol on an Aviv 202 CD spectrometer using a 1-mm pathlength cuvette. Spectra were recorded from 195 to 260 nm at 25 °C, with 1 nm wavelength steps and the measurement at each wavelength being averaged for 30 s. A temperature melt study was carried out by increasing the temperature of the sample from 4 to 90 °C in 1 °C increments, with the ellipticity being monitored at 216 nm. The sample was then incubated at 90 °C for 2 min and cooled from 90 to 4 °C in 1 °C decrements, with the ellipticity being monitored at 216 nm.

Tetranucleotide composition analysis. Tetranucleotide composition analysis can be used to identify core genome signatures to aid in taxonomic assignment, or to differentiate conserved protein-coding regions from those that were horizontally acquired^{44–46}. Tetranucleotide distributions of Nanoarchaeota genomes were determined as previously described⁴³, using a custom Python script. Briefly, sequences were fragmented with a 5-kb sliding window (500-bp overlapping step). Tetranucleotide frequencies were calculated by a zero-order Markov method, which applies odds ratios of observed counts for the 256 unique 4-mers, normalized to their respective mononucleotide frequencies. In order to assess tetranucleotide signatures for DGR regions (~ 2 kb each), while avoiding a compositional bias of flanking sequence, we concatenated DGR cassettes from both OTU1 and OTU2 and fragmented this DGR-specific sequence (~ 21 kb) with a

sliding window as above. In addition, sequences from RT genes and TP genes were separately concatenated and fragmented with a sliding window as above to compare tetranucleotide compositions for the two DGR components.

Dimensionality reduction was performed via non-metric multidimensional scaling on Euclidean distances, using the vegan package in R⁴⁷, and ordination ellipses representing the 95% confidence region were drawn with the 'ordiellipse()' function.

References

- Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C. & DHondt, S. Global distribution of microbial abundance and biomass in seafloor sediment. *Proc. Natl. Acad. Sci. USA* **109**, 16213–16216 (2012).
- Lipp, J., Morono, Y., Inagaki, F. & Hinrichs, K.-U. Significant contribution of Archaea to extant biomass in marine subsurface sediments. *Nature* **454**, 991–994 (2008).
- Valentine, D. L. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat. Rev. Microbiol.* **5**, 316–323 (2007).
- Hoehler, T. M. & Jørgensen, B. B. Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.* **11**, 83–94 (2013).
- Lewin, A. et al. The microbial communities in two apparently physically separated deep subsurface oil reservoirs show extensive DNA sequence similarities. *Environ. Microbiol.* **16**, 545–558 (2014).
- Liu, M. et al. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science* **295**, 2091–2094 (2002).
- Doulatov, S. et al. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481 (2004).
- Medhekar, B. & Miller, J. F. Diversity-generating retroelements. *Curr. Opin. Microbiol.* **10**, 388–395 (2007).
- McMahon, S. A. et al. The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat. Struct. Mol. Biol.* **12**, 886–892 (2005).
- Guo, H. et al. Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Mol. Cell* **31**, 813–823 (2008).
- Le Coq, J. & Ghosh, P. Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. *Proc. Natl. Acad. Sci. USA* **108**, 14649–14653 (2011).
- Rohwer, F. & Vega Thurber, R. Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009).
- Rowlands, T., Baumann, P. & Jackson, S. P. The TATA-binding protein: a general transcription factor in eukaryotes and archaeobacteria. *Science* **264**, 1326–1329 (1994).
- Dwivedi, B., Xue, B., Lundin, D., Edwards, R. A. & Breitbart, M. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evol. Biol.* **13**, 33 (2013).
- Arambula, D. et al. Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. *Proc. Natl. Acad. Sci. USA* **110**, 8212–8217 (2013).
- Schillinger, T., Lisfi, M., Chi, J., Cullum, J. & Zingler, N. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics* **13**, 430 (2012).
- Goldrath, A. W. & Bevan, M. J. Selecting and maintaining a diverse T-cell repertoire. *Nature* **402**, 255–262 (1999).
- Alder, M. N. et al. Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* **310**, 1970–1973 (2005).
- Stokke, R., Roalkvam, I., Lanzen, A., Halfidason, H. & Steen, I. H. Integrated metagenomic and metaproteomic analyses of an ANME-1-dominated community in marine cold seep sediments. *Environ. Microbiol.* **14**, 1333–1346 (2012).
- Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
- Huber, H. et al. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
- Podar, M. et al. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* **8**, 9 (2013).
- Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. USA* **109**, 3962–3966 (2012).
- Simon, D. M. & Zimmerly, S. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res.* **36**, 7219–7229 (2008).
- Ye, Y. Identification of diversity-generating retroelements in human microbiomes. *Int. J. Mol. Sci.* **15**, 14234–14246 (2014).
- Louis-Jeune, C., Andrade-Navarro, M. A. & Perez-Iratxeta, C. Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins* **80**, 374–381 (2012).
- Paull, C. K., Normark, W. R., Ussler, W., Caress, D. W. & Keaten, R. Association among active seafloor deformation, mound formation, and gas hydrate growth and accumulation within the seafloor of the Santa Monica Basin, offshore California. *Mar. Geol.* **250**, 258–275 (2008).

28. Widdel, F. & Bak, F. in: *The Prokaryotes* 2nd edn (eds Balows, A., Trüper, H. G., Dworkin, M., Harder, W. & Schleifer, K.-H.) (Springer, 1992).
29. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
30. Henn, M. R. *et al.* Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE* **5**, e9083 (2010).
31. Schmieder, R., Lim, Y., Rohwer, F. & Edwards, R. TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**, 341 (2010).
32. Hurwitz, B., Deng, L., Poulos, B. & Sullivan, M. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15**, 1428–1440 (2013).
33. Niu, B., Fu, L., Sun, S. & Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*. **11**, 187 (2010).
34. Sun, S. *et al.* Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.* **39**, D546–D551 (2011).
35. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
37. Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* **32**, D45–D49 (2004).
38. Hurwitz, B. L. & Sullivan, M. B. The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE* **8**, e57355 (2013).
39. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371 (2009).
40. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
41. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
42. Kumar, S., Nei, M., Dudley, J. & Tamura, K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform.* **9**, 299–306 (2008).
43. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
44. Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145–158 (2003).
45. Teeling, H., Meyerdieck, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938–947 (2004).
46. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10**, R85 (2009).
47. Oksanen, J. *et al.* Vegan: Community Ecology Package. R package version 1.13-1 <http://vegan.r-forge.r-project.org/> (2008).

Acknowledgements

This research was funded by National Science Foundation grant OCE-1046144 to D.L.V. and National Institutes of Health grant RO1 AI069838 to P.G. and J.F.M.; sequencing was provided through a Gordon and Betty Moore Foundation grant to the Broad Institute. We thank Tanja Woyke for assistance in examining *Nanoarchaeota* sequences from the Microbial Dark Matter project. For assistance with viral metagenome preparation and advice on bioinformatic analyses, we thank Steven Quistad and Rob Edwards. Yanling Wang provided helpful comments on an earlier draft of the manuscript.

Author contributions

B.G.P. performed the sediment incubations and purified viral DNA. B.G.P. and S.C.B. performed preprocessing and annotation of the metagenomic data set. B.G.P., S.C.B., E.C., D.A., S.H., A.S., P.G., J.F.M. and D.L.V. conducted bioinformatic analyses of DGR sequences. S.H. and P.G. expressed and assayed nanoarchaeal target proteins and analysed the resulting data. B.G.P., S.C.B. and D.L.V. wrote the manuscript.

Additional information


Accession codes: Metagenomic sequence reads have been deposited in the NCBI BioSample database with accession code PRJNA47435.DV-ANM1. The ANMV-1 assembled genome sequence has been deposited in the NCBI nucleotide database with the accession code KP703175.

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: J.F.M. is a cofounder, equity holder and chair of the scientific advisory board of AvidBiotics Inc., a biotherapeutics company in San Francisco. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Paul, B. G. *et al.* Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nat. Commun.* **6**:6585 doi: 10.1038/ncomms7585 (2015).

 This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

APPENDIX B. A new topology of the HK97-like fold revealed in *Bordetella* bacteriophage by cryoEM at 3.5 Å resolution



A new topology of the HK97-like fold revealed in *Bordetella* bacteriophage by cryoEM at 3.5 Å resolution

Xing Zhang¹, Huatao Guo², Lei Jin², Elizabeth Czornyj², Asher Hodes²,
Wong H Hui¹, Angela W Nieh², Jeff F Miller^{1,2,3}, Z Hong Zhou^{1,2,3*}

¹California NanoSystems Institute, University of California, Los Angeles, Los Angeles, United States; ²Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, United States; ³Molecular Biology Institute, University of California, Los Angeles, Los Angeles, United States

Abstract Bacteriophage BPP-1 infects and kills *Bordetella* species that cause whooping cough. Its diversity-generating retroelement (DGR) provides a naturally occurring phage-display system, but engineering efforts are hampered without atomic structures. Here, we report a cryo electron microscopy structure of the BPP-1 head at 3.5 Å resolution. Our atomic model shows two of the three protein folds representing major viral lineages: jellyroll for its cement protein (CP) and HK97-like ('Johnson') for its major capsid protein (MCP). Strikingly, the fold topology of MCP is permuted non-circularly from the Johnson fold topology previously seen in viral and cellular proteins. We illustrate that the new topology is likely the only feasible alternative of the old topology. β -sheet augmentation and electrostatic interactions contribute to the formation of non-covalent chainmail in BPP-1, unlike covalent inter-protein linkages of the HK97 chainmail. Despite these complex interactions, the termini of both CP and MCP are ideally positioned for DGR-based phage-display engineering.

DOI: 10.7554/eLife.01299.001

*For correspondence:
Hong.Zhou@UCLA.edu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 17

Received: 27 July 2013

Accepted: 27 October 2013

Published: 17 December 2013

Reviewing editor: Stephen C Harrison, Harvard Medical School, United States

© Copyright Zhang et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

Capsid proteins of non-enveloped viruses fall, so far, into three major structural classes: the β -jellyroll, the HK97 fold, and the fold of dsRNA-virus shell proteins (Bamford et al., 2005; Oksanen et al., 2012). The RNA bacteriophage MS2 subunits have a fourth structure, not yet found in eukaryotic viruses (Valegard et al., 1990). The HK97 fold is present not only in a large number of dsDNA bacteriophage capsids, including the T-phages, lambdoid phages, etc, but also in the major capsid protein of eukaryotic herpesviruses (Bamford et al., 2005). In HK97 itself, the intersubunit contacts are reinforced by a post-assembly covalent linkage—an isopeptide bond between adjacent gp5 subunits, so positioned that the entire capsid is topologically interlinked in a 'chainmail' arrangement (Duda, 1998; Wikoff et al., 2000). In other cases, such as bacteriophage λ , the non-covalent interactions between the subunits are reinforced by an additional 'cement' protein, which binds on the outer surface of the capsid at positions close to those of the isopeptide bonds in HK97 (Lander et al., 2008). Heads of bacteriophage λ defective in this cement protein break down during DNA packaging (Sternberg and Weisberg, 1977; Fuller et al., 2007; Lander et al., 2008).

Bacteriophage BPP-1 is a short-tailed, dsDNA virus and a member of the *Podoviridae* family. It infects and kills *Bordetella* species that cause whooping cough in humans and respiratory diseases in other mammals. It has a T = 71 icosahedral capsid, ~670 Å in diameter. A 7 Å resolution cryoEM structure of BPP-1 showed that its capsid protein has an HK97-like fold, but the shell has an additional protein component (Dai et al., 2010). Lacking information at the time about their genetic identities,

eLife digest Whooping cough is a respiratory illness caused by bacteria in the *Bordetella* genus. Among the general public, *Bordetella* species have become a hot topic in recent years due to the re-emergence of whooping cough in the United States and elsewhere. Scientists, meanwhile have become interested in a virus called BPP-1 that can kill the *Bordetella* species.

BPP-1 is a double-stranded DNA virus, and such viruses have long been of interest to scientists because they are the most abundant organisms on Earth. These viruses are also noteworthy because their shells (also known as capsids) are capable of withstanding the very high pressures (up to about 40 atmospheres) that are created by packing so much DNA into the very small volume inside the capsid.

BPP-1 is of particular interest because it is capable of making large-scale changes to its own DNA in order to adapt to changes in its hosts and environment. Of all the organism that do not contain nuclei within their cells (collectively known as prokaryotes), BPP-1 is the only one that is capable of making such changes to its DNA. However, efforts to exploit the properties of BPP-1 for bioengineering applications have been hampered because its detailed structure is not known. Now Zhang et al. have used cryo electron microscopy to study the structure of BPP-1 at the atomic level.

Most viruses belong to one of three major lineages, with each lineage having a characteristic fold in its capsid proteins. Zhang et al. found that BPP-1 contains two of these folds, which suggests that it is a hybrid of two of these lineages. This is the first time that such a structure has been observed. Moreover, Zhang et al. found that one of the folds has an unusual topology that has not been seen before. The atomic structure reveals how double-stranded DNA viruses use a variety of non-covalent interactions and a type of protein 'chainmail' to form a highly stable capsid that is capable of withstanding very high pressures.

In addition to enabling applications in bioengineering, the new structure might also provide insights into the evolution of prokaryotes.

DOI: 10.7554/eLife.01299.002

these proteins were named according to their structural roles: a major capsid protein (MCP) and a 'cement protein' (CP) that decorates the shell.

From a biotechnology standpoint, BPP-1 has emerged as an attractive phage-display system thanks to its unique and well characterized diversity-generating retro-element (DGR). As the only known source of massive DNA sequence deviation in prokaryotes, DGRs use a unique reverse transcriptase-based mechanism to introduce targeted diversity into protein-coding DNA sequences to accelerate the evolution of adaptive traits (Liu et al., 2002; Guo et al., 2008). As such, BPP-1 is a naturally occurring diversity-generating system and can be engineered to display foreign proteins with adaptive heterologous sequences. An atomic description of the BPP-1 head will enhance bioengineering efforts, reveal non-covalent molecular interactions conducive of stable bacteriophage capsid formation, and clarify evolutionary relationships of BPP-1 MCP and CP with proteins in other viruses.

In this study, we report the three-dimensional (3D) structure of the BPP-1 head at ~3.5 Å resolution determined by single-particle cryoEM and derived an atomic model. Both our structural and structure-based mutagenesis studies reveal major novelties in the BPP-1 MCP and CP structures and their interactions. We also show that the C-termini of both MCP and CP are ideally positioned to display DGR-diversified peptide libraries for protein engineering applications.

Results

Identification of BPP-1 head proteins by mass spectrometry

Genomic analysis indicated that the 42.5 kb BPP-1 genome encodes up to 50 viral proteins (Liu et al., 2004). To identify the genes coding for MCP and CP, we first carried out SDS-PAGE analysis and showed that the two most abundant protein bands have molecular masses of 36.3 kD and 15.2 kD (Figure 1A). Mass spectrometry analysis confirmed that the bands correspond to Bbp17 and Bbp16, respectively (Figure 1B). The theoretical molecular masses of Bbp17 and Bbp16 (i.e., 36417.2D and 14458.3D) match their apparent molecular masses by SDS-PAGE, suggesting that the two capsid proteins in the mature bacteriophage particles are not cleaved (Figure 1A). Finally, to verify that genes

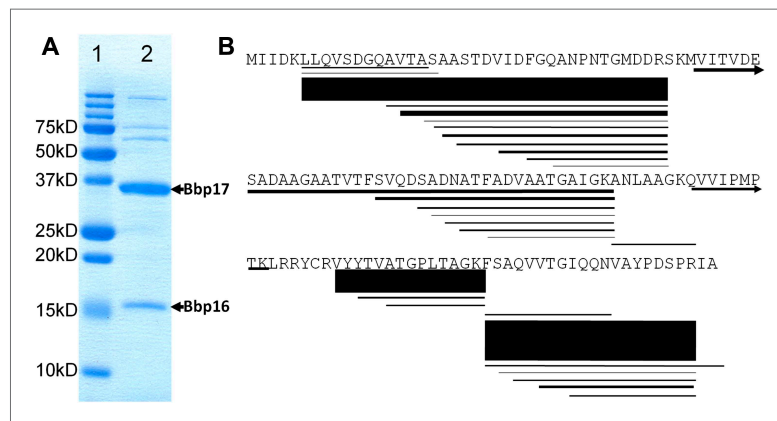


Figure 1. Identification of BPP-1 capsid proteins. **(A)** SDS-PAGE of BPP-1 virion proteins stained with Coomassie blue; lane 1: molecular mass standards; lane 2: BPP-1 virion proteins. The two most abundant proteins of the BPP-1 virion are identified by mass spectrometry **(B)** to be Bbp17 (MCP) and Bbp16 (CP) (indicated by arrows). **(B)** Mass spectrometry result of Bbp16 (CP). The sequence is shown with individual peptides identified by mass spectrometry drawn as lines below their corresponding sequences, with line thickness and darkness representing relative abundance in the mass spectrometry profiles (thicker lines mean more abundant). Arrows indicate the two peptide fragments that run past the end of the rows.

DOI: 10.7554/eLife.01299.003

bbp16 and *bbp17* encode components essential for forming infectious phage particles, we constructed in-frame deletions of the *bbp16* and *bbp17* genes. As expected, no infectious phage particles were produced from BPP-1 Δ *brt* lysogens carrying deletions in *bbp16* or *bbp17*, confirming that both MCP (Bbp17) and CP (Bbp16) are required for phage production. For the sake of clarity and ease of comparison with structural homologs of other viruses, we will continue to use CP and MCP to refer to Bbp16 and Bbp17 of BPP-1, respectively. As shown below, these assignments are directly verified through atomic modeling in which side-chain structures of amino acid residues in sequences match side-chain densities visualized in the cryoEM map (Figure 2).

CryoEM structure of the BPP-1 head

We reconstructed the 3D structure of the BPP-1 icosahedral head by single-particle cryoEM (Figure 2, Video 1). The quality of MCP and CP densities was further improved by additional averaging of seven CP or six hexameric MCP subunits (except for the pentameric MCP) in the asymmetric unit (Zhang *et al.*, 2010a). Based on the reference-based Fourier shell correlation criterion (FSC = 0.143), the resolution of the capsid is 3.58 Å (Figure 2C) (Rosenthal and Henderson, 2003). Consistently, the R-factors of atomic models are better than 0.5 at the zone of 1/3.5 Å (for capsid) or 1/3.4 Å (for averaged CP and MCP) (Figure 2C; Table 1), which correspond to a Fourier shell correlation (FSC) coefficient greater than 0.143 (Wolf *et al.*, 2010). Therefore, the resolution of the capsid and averaged densities were estimated to be ~3.5 Å and ~3.4 Å, respectively. This assessment of resolution is also consistent with clearly visible side chain densities of bulky amino acid residues, such as arginine and phenylalanine (Figures 2D and 4E–F; Videos 2–4).

Arranged on a T = 7 icosahedral shell, MCP and CP each have seven copies or conformers in each asymmetric unit. MCP alone forms a complete icosahedral shell without any noticeable gaps. CP forms dimers bound to the underlying MCP shell at local and icosahedral twofold axes, making the maximum diameter of the BPP-1 (670 Å) bigger than that of the HK97 capsid (659 Å). When CP dimers are computationally removed, the overall size and architecture of the BPP-1 capsid is almost identical to that of the HK97 capsid, and the backbone model of the HK97 capsid is nearly super-imposable with the BPP-1 density map with minor mismatches. Based on the averaged density maps of CP and MCP, we built initial atomic models of CP and MCP with Coot (Emsley and Cowtan, 2004), and refined

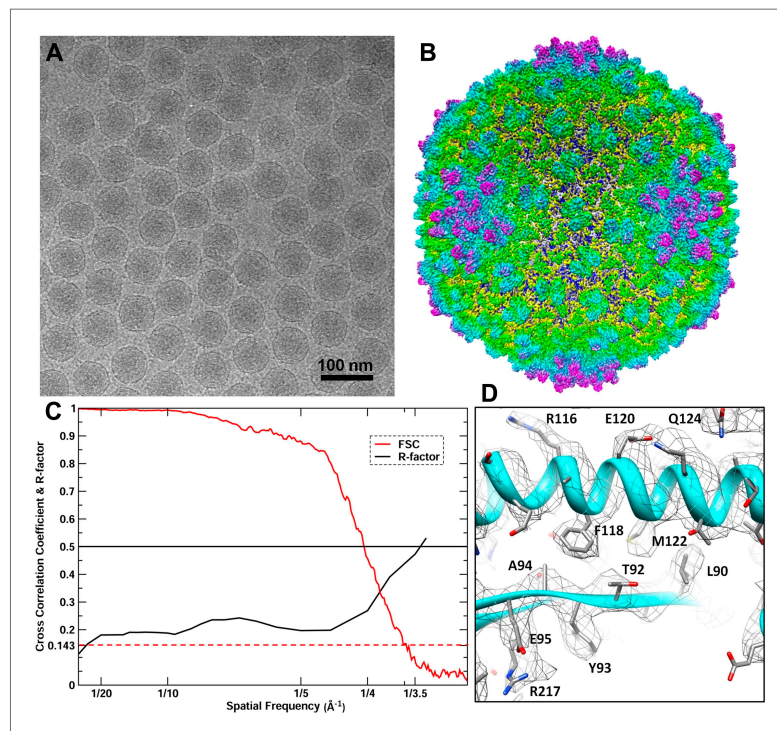


Figure 2. CryoEM reconstruction of the BPP-1 head at 3.5 Å resolution. (A) Representative cryoEM image (defocus=1.6 μm) of the BPP-1. (B) CryoEM density of the BPP-1 head shown as radially-colored surface representation. See also [Video 1](#). (C) R-factors (red) and Fourier shell correlation coefficients (FSC) (black) as a function of spatial frequency between maps from half datasets. (D) Close-up view of a local region of MCP, with densities of many amino acid side chains clearly resolved in both the helix and the loop. The atomic model is shown as ribbons and sticks with amino acid residues labeled.

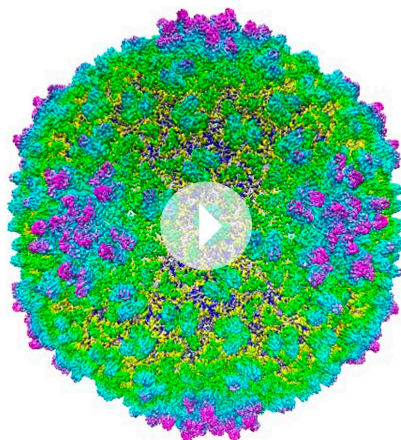
DOI: [10.7554/eLife.01299.004](https://doi.org/10.7554/eLife.01299.004)

them with Phenix ([Adams et al., 2010](#)). Models of the seven MCP and CP monomers within the asymmetric unit were subsequently obtained by adjusting the initial MCP and CP models with Coot and refined with Phenix, with overall R/R-free factors of 0.26/0.27 at 3.5 Å resolution ([Table 1](#)) ('Materials and methods').

Cement protein Bbp16 has the jellyroll fold

We traced 139 (Met1 to Ile139) of the 140 amino acid residues of the cement protein Bbp16 in the cryoEM density map. CP contains two β-sheets each consisting of four anti-parallel strands, and a ~36-Å long extension (Thr125 to Ala140) with the C-terminus exposed on the external surface ([Figure 3](#); [Video 2](#)). The backbones of the seven CP subunits in the asymmetric unit, including their C-terminal extensions, are nearly identical, with an RMSD of 0.4–0.66 Å when superimposed ([Figure 3D](#)).

The topological organization of strands in the two β sheets (i.e., BIDG and CHEF; [Figure 3B](#)) of CP is identical to that of the jellyroll motif. This structural motif was first discovered in spherical RNA viruses ([Harrison et al., 1978](#); [Abad-Zapatero et al., 1980](#); [Hogle et al., 1985](#); [Rossmann et al., 1985](#); [Chelvanayagam et al., 1992](#)), and subsequently found widely in other DNA viruses, such as φX174, bacteriophage PRD1 and human adenovirus ([McKenna et al., 1992](#); [Abrescia et al., 2004](#); [Zubieta et al., 2005](#); [Liu et al., 2010](#); [Krupovic and Bamford, 2011](#)). Unlike its role as the major capsid protein in the above-mentioned viruses, the jellyroll motif in BPP-1 forms an auxiliary protein to



Video 1. Shaded surface view of the cryoEM density of the BPP-1 capsid at 3.5 Å resolution. The map is color-coded according to the radius. Related to **Figure 2**. DOI: 10.7554/eLife.01299.005

Video 3). Although little sequence identity (CLUSTALW score = 5) was detected between BPP-1 MCP and HK97 gp5 proteins with ClustalW (Thompson *et al.*, 1994), the overall architecture of the MCP resembles that of HK97 gp5 (Wikoff *et al.*, 2000) (Figure 4—figure supplement 1A), with an axial (A) domain, a peripheral (P) domain and an extended loop (E-loop) (Figure 4C). Both the subunit organization and domain orientations of the seven MCPs in the BPP-1 capsid are also identical to those of the corresponding gp5 subunits in the HK97 capsid. The A-domain contains a 5-stranded β sheet flanked by two helices on one side and a C-terminal loop on the other. The P-domain contains a characteristic long, 7-turn helix sandwiched by a three-stranded β sheet and the N-terminal loop. The E-loop contains a single amino acid sequence segment folded into an extended hairpin loop projecting from a 2-stranded β sheet. Notably, both N- and C-termini of MCP are exposed on the external surface of the capsid, thus are accessible for tethering peptides in phage-display applications.

The structures of the seven MCP conformers in the asymmetric unit are nearly identical with some minor differences (Figure 4—figure supplement 1B), and RMSDs of their backbones (residues from 31–60 to 81–331) are 0.55–0.98 Å among six hexameric MCPs and 1.2–1.8 Å between

stabilize the viral capsid made by proteins of another fold. Notably, in some dsRNA viruses, the jellyroll motif has been adopted as a domain in a stabilizer/adaptor protein that forms trimers located at an intermediate layer of viral capsids (Grimes *et al.*, 1998; Mathieu *et al.*, 2001; Liemann *et al.*, 2002; Zhang *et al.*, 2010a). In BPP-1, CP subunits form a dimer through β -sheet augmentation between the F-strands of two CHEF sheets, forming an 8 stranded β -sheet visible on the external surface (Figure 3C). Although the BIDG sheet of the jellyroll motif faces but do not interact with the underneath MCP shell, instead, as described in detail below, CP interacts with the MCP shell mainly through its N- and C-termini as well as a linking loop between its β -strands B and C.

Structure of the major capsid protein Bbp17 reveals a new topology of the Johnson fold

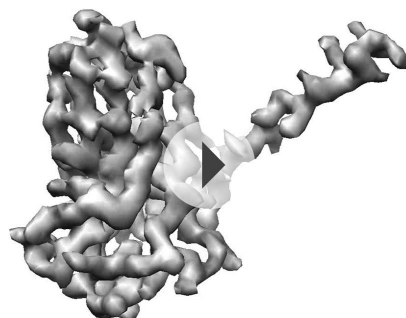
We traced 327 amino acid residues (Ser5 to Val331) of the total 331 residues of the major capsid protein Bbp17 in the cryoEM density map (Figure 4;

Table 1. Statistics of atomic model refinement with Phenix

	CP (Bbp16)	MCP (Bbp17)	Asymmetric unit (7CPs+7MCPs)
Residues resolved	1–140	7–331	1–140 (CP); 5–331 (MCP)
Resolution (Å)	3.4	3.4	3.5
R_{work} (overall: 40–3.4 Å)	0.28	0.27 (0.31*)	0.26
R_{free} (overall: 40–3.4 Å)	0.28	0.28 (0.30*)	0.26
R_{work} (best resolution zone)	0.47 (1/3.4 Å)	0.50 (0.52*) (1/3.4 Å)	0.51 (1/3.5 Å)
R_{free} (best resolution zone)	0.45 (1/3.4 Å)	0.45 (0.50*) (1/3.4 Å)	0.47 (1/3.5 Å)
Ramachandran plot values			
Most favored (%)	85.5	86.7	88.0
Generously allowed (%)	12.3	11.8	10.3
Disallowed regions (%)	2.2	1.5	1.7

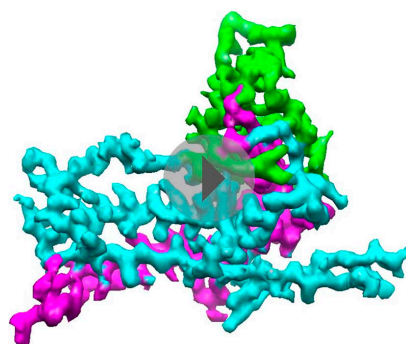
*R-factors of the interchanged model by forcing the BPP-1 MCP to trace the HK97 topology.

DOI: 10.7554/eLife.01299.006



Video 2. Shaded surface view of the cryoEM density of the averaged CP at 3.4 Å resolution (gray) superimposed with the atomic model CP (ribbon and sticks). Related to [Figure 3](#).

DOI: [10.7554/eLife.01299.007](https://doi.org/10.7554/eLife.01299.007)



Video 3. Shaded surface view of the cryoEM density of the averaged MCP at 3.4 Å resolution. The three structural elements are color coded (cyan, purple and green) according to the structural elements of the Johnson fold as illustrated in [Figure 4](#). Related to [Figure 4](#).

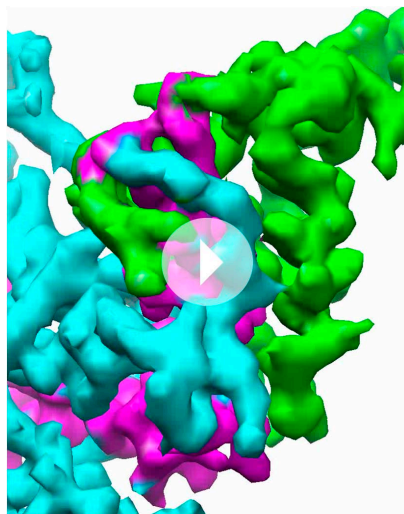
DOI: [10.7554/eLife.01299.008](https://doi.org/10.7554/eLife.01299.008)

the pentameric and the hexameric MCPs. The minor differences among the different MCP conformers include: (1) Each N-terminal portion (Ser5–Glu30) of the seven MCPs adopts a slightly different conformation, and backbone RMSDs of this segment are 1.4–5.0 Å among seven MCP copies; (2) The E-loops of the seven MCPs have backbone RMSDs of 0.8–8.3 Å, larger than other portions of MCPs, possibly resulting from different local shell curvatures at the regions of the E-loops ([Figure 4E](#)); (3) Although the C-terminal proximal loop (Ala295–Val309) of every hexameric MCP is well resolved, the corresponding segment in the pentameric MCP is not resolved, suggesting disordered conformation of this segment in pentons possibly due to steric hindrance at the central channel of the MCP pentamer ([Figure 4—figure supplement 2](#)).

Despite the similarities in the architectural appearances of BPP-1 MCP and HK97 gp5, their atomic structures differ in four significant ways. Firstly, the conformations of the N-terminal loops are different ([Figure 4—figure supplements 1, 2C](#)). In BPP-1 MCP, the N-terminal peptide extends radially to interact with a CP by augmenting its CHEF β -sheet. In HK97 gp5, the N-terminus extends circumferentially and interacts with three adjacent gp5 subunits, augmenting a β -sheet in one of them. Secondly, a C-terminal loop (Lys293–Val309) of the pentameric MCP in BPP-1 is disordered while the corresponding loop (Arg294–Thr304) in the HK97 pentameric gp5 is ordered ([Figure 4—figure supplement 2B,C](#)) ([Wikoff et al., 2000](#)). Thirdly, the electrostatic potential properties of the two proteins are strikingly different. Finally, and perhaps most interestingly, the folding topologies of BPP-1 MCP and HK97 gp5 are different, which, as described below, provides the first opportunity to explore possible alternative ways to build the highly conserved HK97-like folds ([Figure 5](#), [Figure 5—figure supplements 1 and 2](#); [Video 4](#)).

Two topologies to build the HK97-like ('Johnson') fold and structure-based mutagenesis

As remarked above, despite of the similar architectural appearance (or fold) of BPP-1 Bbp17 and HK97 gp5, the peptide traces of these two proteins follow different folding topologies ([Figure 5A–B](#), [Figure 5—figure supplements 1 and 2](#)). A careful comparison of the two structures has led us to identify three structural elements of the canonical HK97-like, or 'Johnson' fold: N-, β - and α -elements (marked cyan, purple and green in [Figures 4A–D and 5A–C](#)). These three structural elements join together through a central 5-stranded (F, E, G, K, L in [Figure 5A–C](#), in an up-up-down-up-up topology) β -sheet, flanked by two short helices ([Figure 5A–C](#), [Figure 5—figure supplement 2](#)). The β -element (purple) consists of two anti-parallel (i.e., G, K; down-up) β -strands forming a hairpin located at the middle of the Johnson fold. The α -element (green) contains two parallel (i.e., F, E; up-up) β -strands and the two short helices. The N-element (cyan) contributes the last (i.e., L; up) β -strand to the Johnson



Video 4. CryoEM density of MCP around the positions of permutation. The three structural elements are color coded (cyan, purple, and green) according to the structural elements of the Johnson fold as illustrated in [Figure 4](#). Related to [Figure 4](#). DOI: [10.7554/eLife.01299.009](https://doi.org/10.7554/eLife.01299.009)

fold and also contains the two other characteristic secondary structures of the Johnson fold: a long, kinked ‘spine’ α -helix and the extended loop (E-loop).

In BPP-1 MCP, the N-, β -, and α -elements consist of residues 1–168, 169–241, and 242–331, respectively ([Figure 4A](#)). The orders of these three structural elements in BPP-1 and HK97 are N- β - α (‘BPP topology’, [Figures 4A–C and 5A](#)), N- α - β (‘HK97 topology’, [Figure 5B](#)), respectively. To verify this, we swapped the α and β structural elements in our de novo BPP-1 MCP model to create an ‘interchanged model’ that matches the HK97 topology. This ‘interchanged model’ was then refined with Phenix for five cycles. Despite this refinement effort, the final ‘interchanged model’ does not agree with our experimental cryoEM density ([Figure 4—figure supplement 3](#)). In particular, many side chains in the ‘interchanged model’ do not match those resolved in the cryoEM density map ([Figure 4—figure supplement 3B, C](#)). This verification provides additional support to the BPP topology established by our de novo modeling approach.

Interchange of two of the structural elements can lead to proteins with different topologies of the Johnson fold ([Figure 5C](#)). From a pure mathematical point of view, permuting three structural elements can give rise to a total of six different

topologies (i.e., $3! = 6$), considering that sequence polarity (N- to C-termini) of the structural elements must be preserved. However, the N-element (cyan) cannot participate in this permutation due to the remote disposition (~ 42 Å) of its N-terminus with respect to the C-termini of other structural elements, and thus the total possible topologies is reduced to only 2 (i.e., $2! = 2$) and the permutation is non-circular ([Figure 5—figure supplements 1A–B, 2G–H](#)). Remarkably, the newly discovered BPP topology presented here is therefore the second topology predicted from the above rule ([Figure 5A](#)), the first being the topology discovered in HK97 gp5 ([Wikoff et al., 2000](#)) ([Figure 5B](#)) and subsequently seen in many other viruses, as well as in bacteria and archaea cells ([Akita et al., 2007](#); [Sutter et al., 2008](#)). Insertions into any of the three structural elements can give rise to more complex architectures without increasing the number of folding topologies of the Johnson fold, as would be expected for larger viral particles such as the herpesvirus capsid.

To test whether a functional protein can be produced from BPP-1 MCP by permuting it to the HK97 topology without introducing other changes, we genetically interchanged the primary order of the β - and α -elements in the BPP-1 MCP gene *bbp17* ([Figure 5—figure supplement 1C](#)), and made two slightly different constructs ([Figure 5—figure supplement 1C](#)) (i.e., PM1 and PM2, ‘Materials and methods’). These two engineered proteins were expressed successfully, as indicated by Western blot analysis ([Figure 5D](#)). However, no plaque formation was detected for the PM1 and PM2 lysates on RB50 cells transformed with either the wt *bbp17*, the PM1 or the PM2 construct ([Figure 5E](#)), suggesting that the PM1 and PM2 gene products are not functional.

Interactions among capsid proteins

The BPP-1 capsid has a chainmail structure similar to that of HK97 capsid ([Figure 6A–B](#); [Video 5](#)). Instead of covalent bonding as in HK97, the BPP-1 chainmail is stabilized by non-covalent interactions, such as: (1) strong electrostatic interactions between adjacent MCPs ([Figure 6](#)), and (2) additional CP-MCP interactions contributed by CPs ([Figure 7](#)).

Electrostatic interactions between MCPs of BPP-1 are stronger than those between gp5 subunits in the HK97 capsid ([Figure 6D–H](#), [Figure 6—figure supplements 1](#)) for two reasons. Firstly, in the MCP

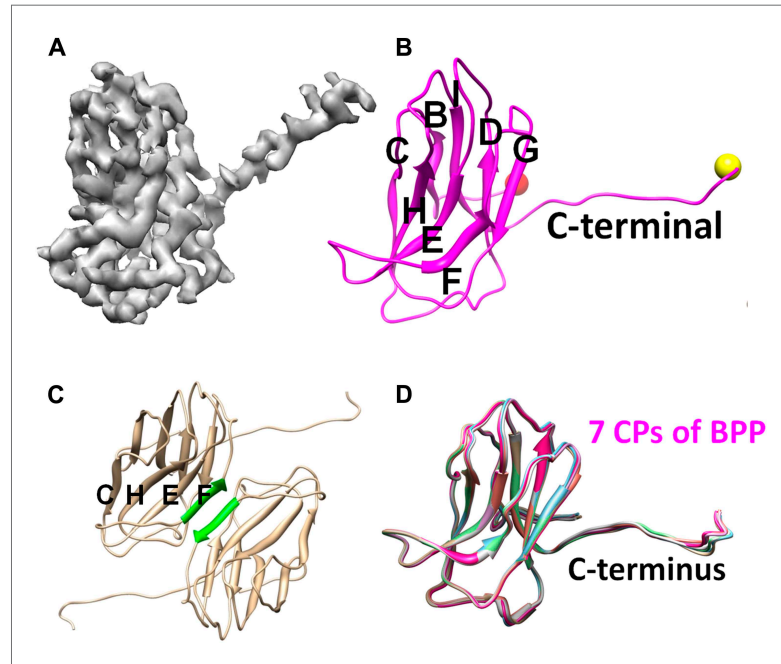


Figure 3. Structure of CP. (A) CryoEM density map of CP (3.4 Å resolution, average of all seven conformers in an asymmetric unit). See also [Video 2](#). (B) Ribbon model of CP, showing its jellyroll fold. Eight β -strands (B, C, D, E, F, G, H, I) fold into the two characteristic β -sheets (BIDG and CHEF), forming a 'jellyroll' (Harrison et al., 1978; Abad-Zapatero et al., 1980; Hogle et al., 1985; Rossmann et al., 1985). The N- and C-termini are marked by red and yellow balls, respectively. (C) The two F strands (green) of two neighboring CP monomers form hydrogen bonds in an antiparallel fashion, creating an augmented, 8-stranded β -sheet and a CP dimer. (D) Ribbon diagrams of the atomic models of the seven CP conformers of BPP-1.

DOI: [10.7554/eLife.01299.010](https://doi.org/10.7554/eLife.01299.010)

hexamer, the electrostatic interactions between adjacent MCPs that appears to be stronger than those between HK97 hexameric gp5 proteins (Figure 6—figure supplement 1A–B). Secondly, the interface between the MCP E-loop and an adjacent MCP protein at each local threefold region contains complementary electrostatic interactions (Figure 6D–E). Specifically, the E-loop is mainly positive charged and its MCP interface is mainly negative charged, and the interaction at this interface appears to be stronger than that of HK97 gp5 in the same interface. Interestingly, the electrostatic properties of BPP-1 MCP and Hk97 gp5 are opposite in this interface (Figure 6—figure supplements 1A–B). The stronger electrostatic interaction between the E-loops and the local threefold region of the BPP-1 MCP may be the reason for the absence (Figure 8) of the salt bridge found between HK97 gp5 proteins at the local threefold interface that is critical for the assembly, stability and maturation of the HK97 capsid (Gertsman et al., 2010).

In addition, the BPP-1 head is further stabilized by interactions between CP and MCP. Complementary electrostatic potential surfaces were also identified at the interface between CP and MCP (Figure 7A–C). Each CP interacts extensively with five MCP subunits underneath, mainly involving three loops: (a) N-terminal loop, (2) C-terminal loop, and (3) the linking loop between strands C and D of the jellyroll motif (Figure 7D–F). Firstly, the N-terminal loop of CP interacts with two MCPs: the N-terminal loop of first MCP (green) and the E-loop of the second MCP (yellow) (Figure 7D–E). Secondly, the C-terminal loop of CP interacts with two MCPs: a β -strand of the β -element of the first MCP (orange) and both N- and C-terminal loops of a second MCP (green) (Figure 7D–E). Thirdly, the

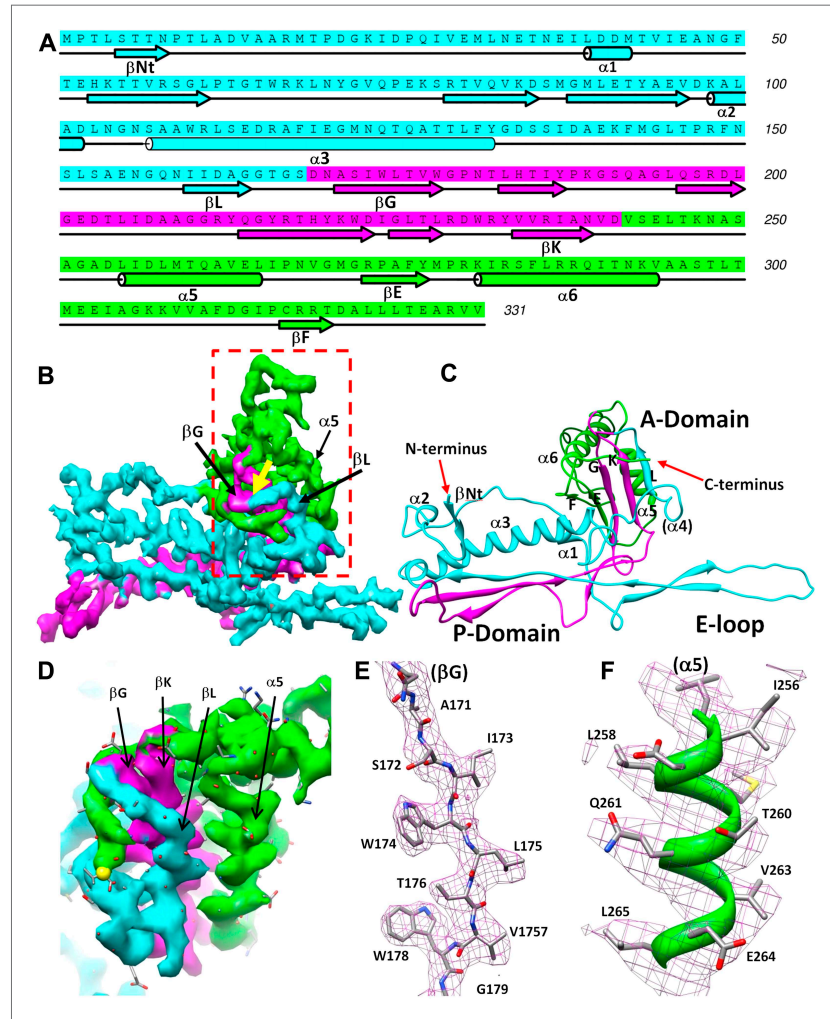


Figure 4. Structure of MCP. (A) Sequence and secondary structure assignment of MCP. α -helices are marked by cylinders, β -strands by arrows, and loops by thin lines. The three structural elements of the Johnson fold, including N-, β - and α -elements, are shown in cyan, purple and green, respectively. (B) CryoEM density map of MCP (3.4 Å resolution, average of the six hexon MCP subunits in an asymmetric unit) shown as shaded surface using the same color coding of (A). (C) Ribbon diagram of the MCP atomic model with the three structural elements of the Johnson fold colored as in (A). (D) BPP-1 MCP density within the dash-box drawn in (B), showing the positions of the permutation (indicated by a yellow arrow) through which the N-element (cyan) is connected to the β -element (purple), instead of to helix α 5 (green, far away), as in HK97 gp5. (E) CryoEM density (mesh) of the β G strand in (D) superimposed with its atomic model (sticks). (F) CryoEM density (mesh) of the α 5 helix in (D) superimposed with its atomic model (ribbon with sticks).

DOI: [10.7554/eLife.01299.011](https://doi.org/10.7554/eLife.01299.011)

Figure 4. Continued on next page

Figure 4. Continued

The following figure supplements are available for figure 4:

Figure supplement 1. Structural comparisons between one hexameric MCP of BPP-1 and gp5 of HK97.

DOI: [10.7554/eLife.01299.012](https://doi.org/10.7554/eLife.01299.012)

Figure supplement 2. Differences between BPP-1 pentameric MCP and HK97 gp5.

DOI: [10.7554/eLife.01299.013](https://doi.org/10.7554/eLife.01299.013)

Figure supplement 3. Incorrect MCP model obtained by enforcing the HK97 topology into the BPP-1 MCP cryoEM density, followed by five cycles of model refinement ('interchanged model').

DOI: [10.7554/eLife.01299.014](https://doi.org/10.7554/eLife.01299.014)

linking loop of CP interacts with three MCPs: two β -strands of the β -element of the first MCP (blue), the N-terminal loop of a second MCP (green) and a short loop (Asn48–Glu52) of the third MCP (orange) (Figure 7D–E). Fourthly, β -strand C of CP jellyroll interacts with the N-terminal loop of an MCP, augmenting the CHEF β -sheet of CP (Figure 9C). Finally, two CP monomers form a dimer through anti-parallel interaction between their F-strands of jellyrolls, forming an augmented, 10-stranded β sheet and further stabilizing the capsid (Figures 3C, 7D and 9C).

Discussion

We found that BPP-1 CP and MCP adopt two of the three characteristic folds of major virus lineages, jellyroll, and HK-97/Johnson. Moreover, the topology of the Johnson fold in BPP-1 MCP is non-circularly permuted as compared to the previously known topology of the Johnson fold. Until now, the Johnson fold has been observed at atomic detail in the *Siphoviridae* and *Myoviridae* families of the three tailed-bacteriophage families, exemplified by gp5 of HK97 (*Siphoviridae*) (Wikoff et al., 2000) and gp24 of T4 (*Myoviridae*) (Fokine et al., 2005). Our atomic model derived from the 3.5 Å cryoEM map of BPP-1 represents the first atomic model for a virus in the third tailed-bacteriophage family, the *Podoviridae*. Indeed, BPP-1 represents the first among all viruses shown to contain both the jellyroll and HK97-like folds. In the virosphere, the conventional wisdom is that viruses have evolved from a few distinctive lineages separately. Three major lineages have been proposed based on three highly distinctive structures: jellyroll-like (i.e., jellyroll lineage, include single- and double-jellyroll), HK97-like (i.e., HK97 lineage) and BTV-like (i.e., BTV lineage), and members of each lineage contain MCPs with conserved folds originating from a common ancestor (Rossmann and Johnson, 1989; Bamford et al., 2005; Abrescia et al., 2012). Our result suggests that BPP-1 is likely a hybrid virus that has adopted the characteristic structures from both the jellyroll and HK97 lineages.

The head of bacteriophage Epsilon15 also contains two capsid proteins (gp7 and gp10) located at similar positions as CP (Bbp16) and MCP (Bbp17) in BPP-1. In particular, BPP-1 Bbp17 (MCP) shares ~46% sequence identity with gp7 of Epsilon15 phage, indicating that the two proteins are likely to have nearly identical structures at least at the level of protein backbones. Indeed, although initially misinterpreted to have a different topology (Figure 5—figure supplement 2H–I) due to limited resolution (Jiang et al., 2008), the model of Epsilon15 gp7 was recently updated on the basis of an improved 4.5 Å resolution map (Baker et al., 2013) to bear the same BPP topology as we initially reported (Zhang et al., 2012) and detailed further in this study.

As described in the results above, the BPP topology and HK97 topology represent the only two feasible topologies of the Johnson fold (Figure 5). To date, all other known atomic structures of proteins with the Johnson fold adopt the HK97 topology, including gp5 of HK97 in *Siphoviridae* (Wikoff et al., 2000), gp24 of T4 in *Myoviridae* (Fokine et al., 2005), a 39kD spherical particle-forming protein in archaea (Akita et al., 2007) and an encapsulin protein in bacteria (Sutter et al., 2008) (Figure 5—figure supplement 2A–E). In these two topologies, the primary positions of their β - and α -elements, are swapped in a non-circularly permuted fashion (Figure 5A–C, Figure 5—figure supplement 1), which rarely occurs in nature (Vogel and Morea, 2006). This, combined with the fact that no recognizable sequence similarity can be identified between the BPP-1 MCP and other HK97-like MCPs (CLUSTALW scores: ~5), obscures the origin of the BPP topology.

From an engineering stand point, it is fortunate that both termini of MCP and the C-terminus of CP are exposed on the external surface of the capsid, ideally positioned for peptide display. Indeed, we observed infectious phage particle formation when either CP or MCP was fused to a peptide of 22 amino acids at their C-terminus (HG and JFM, unpublished observation). As BPP-1 DGRs diversify

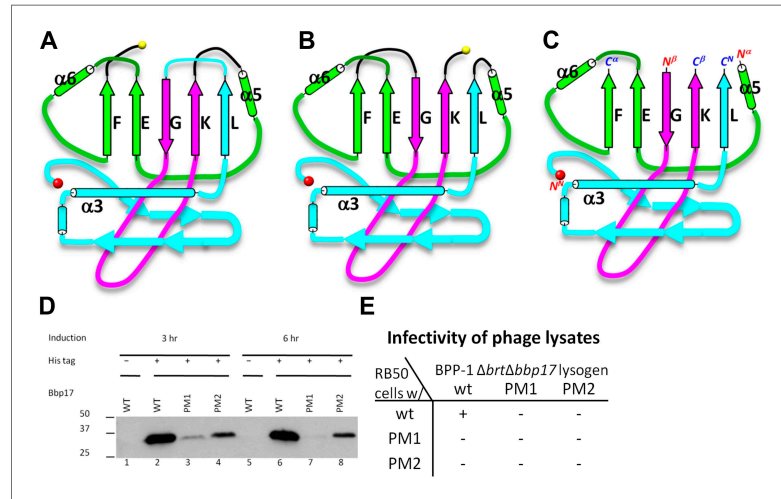


Figure 5. Two different topologies of the Johnson fold and structure-based mutagenesis of MCP. **(A)** Diagram of the BPP topology of the Johnson fold: N-, β -, α -elements. **(B)** Diagram of the HK97 topology of the Johnson fold: N-, α -, β -elements. **(C)** Diagram of the three structural elements of the Johnson fold with free N- and C-ends. **(D)** Expression of BPP-1 MCP mutants with the β - and α -elements swapped, thus adopting the HK97 topology. Wild type (WT) and mutant (PM1 and PM2) BPP-1 MCPs with a 6xhistidine tag at the C-terminus were induced for expression from an *fhaB* promoter in RB50 cells for 3 (lanes 1–4) and 6 (lanes 5–8) hr, respectively. The expressions levels were determined by Western blot with a mouse anti-6xhistidine monoclonal antibody. Lanes 1 and 5 are negative controls with wild type *Bbp17* that does not contain a 6xhistidine tag. **(E)** Infectivity of phage lysates as measured by their ability to form plaques on transformed RB50 cells. '+': plaque observed; '-': no plaque observed. The color scheme in **(A–C)** is the same as in **Figure 4A–D**.

DOI: [10.7554/eLife.01299.015](https://doi.org/10.7554/eLife.01299.015)

The following figure supplements are available for figure 5:

Figure supplement 1. Diagram of non-circular permutation of the three structural elements of the Johnson fold in BPP-1 MCP and HK97 gp5.

DOI: [10.7554/eLife.01299.016](https://doi.org/10.7554/eLife.01299.016)

Figure supplement 2. Johnson fold in HK97-like proteins.

DOI: [10.7554/eLife.01299.017](https://doi.org/10.7554/eLife.01299.017)

DNA sequences in vivo, this suggests they may be engineered to display heterologous sequences on the phage capsid by fusion to the termini of CP or MCP. For DGR-based phage-display, the diversity of the DNA library is not limited by the efficiency of DNA transformation, and both the library construction and optimization can occur entirely inside bacterial cells without the need for in vitro manipulation. Moreover, the selected phage particles can be easily re-amplified for iterative selection since the tail, not the engineered head, is involved in phage-receptor recognition. The atomic model of the BPP-1 head presented in this study provides a road map forward to harness the great potentials afforded by the unique properties of BPP-1 for phage-display engineering.

Materials and methods

Production and purification of BPP-1

500 ml of LB medium was inoculated with a single colony of BPP-1 lysogen. The cultures were incubated at 37°C on a rotary shaker until log phase when phage production was induced by adding mitomycin C to a final concentration of 2 mg/l. After 3 hr of induction, CHCl_3 was added with shaking to facilitate cell lysis and phage release. Cellular debris was removed by centrifugation at 5000×g. Phage particles were then precipitated using 10% PEG8000/500 mM NaCl, pelleted by centrifugation, and resuspended in 50 mM Tris-HCl, 250 mM NaCl, 1 mM MgCl_2 , pH7.5. The resuspended phage

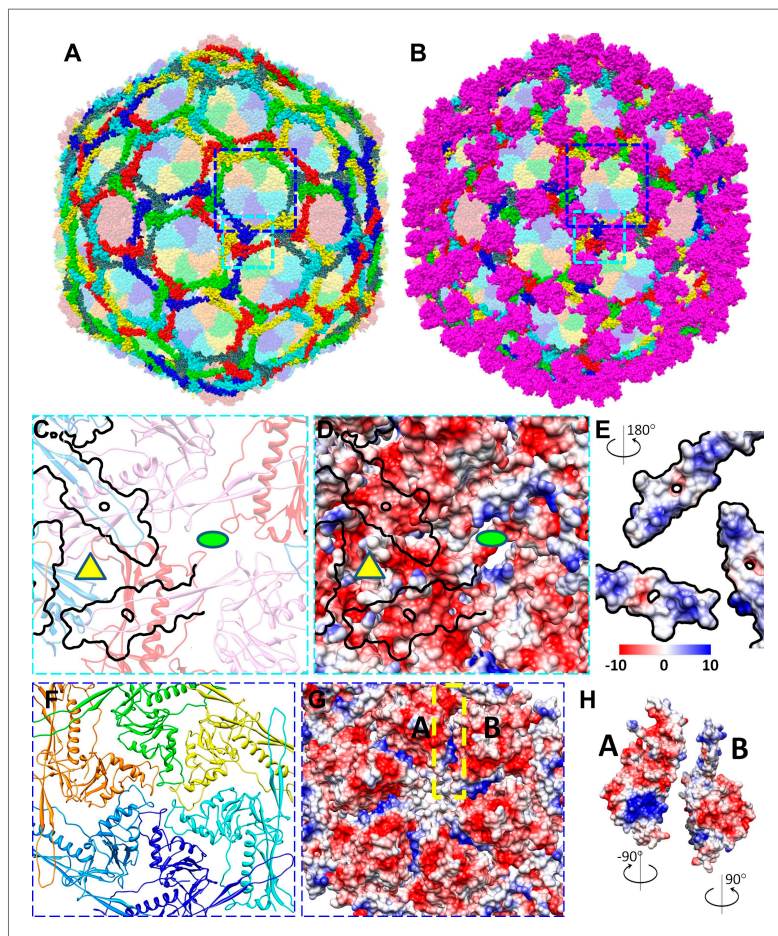


Figure 6. Non-covalent chainmail of the BPP-1 head. See also [Video 5](#). (A) Chainmail network formed by BPP-1 MCPs. The P-domain and E-loop of MCP in neighboring capsomers join head to tail to form rings (bright colors), which concatenate to form non-covalent chainmail. The P-domains and E-loops contributing to the formation of the same ring are shown in the same color. Other domains of MCP are dimmed. (B) Same view as in (A) but with CP dimers (purple) also shown. (C–E) Inter-capsomeric MCP-MCP interactions. The close-up view of all MCPs within the region within the cyan box in (A) illustrates the interaction interfaces (outlined by black lines) between the overlying E-loops of three MCP monomers and the underlying domains of other MCPs. These MCP monomers are shown as ribbons in different colors and belong to different hexon (or penton) capsomers. The local threefold and twofold axes are denoted by a yellow triangle and a green ellipse, respectively. Complementary electrostatic potentials are evident at the interaction interfaces, shown separately to reveal the surfaces of the underlying MCPs (D) and the 180°-rotated overlying E-loops (E), respectively. (F–H) Intra-capsomeric MCP-MCP interactions. The region within the blue box of (A) contains a hexon with its six MCP monomers shown either as ribbons in different colors (F) or as electrostatic potential surfaces (F). Adjacent MCP subunits within the hexon share one interaction interface (e.g., the yellow dashed rectangle for subunits A and B in F) and their complementary electrostatic potential surfaces are evident in their rotated views (H). The electrostatic potential scale is shown in the color bar in (E).
DOI: [10.7554/eLife.01299.018](https://doi.org/10.7554/eLife.01299.018)

Figure 6. Continued on next page

Figure 6. Continued

The following figure supplements are available for figure 6:

Figure supplement 1. Inter-capsomer interactions in HK97 capsid.

DOI: [10.7554/eLife.01299.019](https://doi.org/10.7554/eLife.01299.019)

particles were further purified by 15–45% (buffered with 50 mM Tris-HCl, 250 mM NaCl, 1 mM MgCl₂, pH7.5) sucrose gradient ultracentrifugation at 35,000 rpm for 90 min using an SW41 rotor in a Beckman L8-80M ultracentrifuge. The phage band was visualized by illuminating the gradient tube and carefully collected by side puncture. Purified phage was buffer-exchanged and concentrated in 50 mM Tris-HCl, 250 mM NaCl, 1 mM MgCl₂, pH7.5 by ultra-filtration with a MW cutoff of 100 kD (Millipore, Billerica, MA). The final concentration of BPP-1 was ~10¹⁴ pfu/ml.

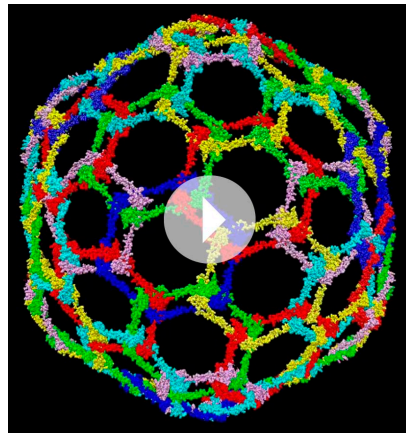
Protein gel electrophoresis and mass spectrometry

To identify capsid proteins using SDS–PAGE and mass spectrometry, purified BPP-1 particles were incubated in 1× SDS sample buffer (0.0625 M Tris-HCl, 1.25% SDS, 5% glycerol and 0.02% bromophenol blue, pH 6.8) at 95°C for 5 min before loading onto the gel. SDS–PAGE was performed using a discontinuous gel (4% polyacrylamide stacking region and 15% polyacrylamide resolving gel) prepared with a discontinuous gel system (Bio-Rad Laboratories, Hercules, CA). Protein bands were visualized by Coomassie Blue staining (Bio-Rad Laboratories). The two major bands (**Figure 1A**) were manually excised, digested in-gel with sequencing-grade modified trypsin (Promega, Madison, WI), and analyzed using Matrix-assisted laser desorption/ionisation-time of flight mass spectrometry (MALDI-TOF MS). Scaffold (version.3.00.04, Proteome Software Inc., Portland, OR) was used to validate MS/MS-based peptide and protein identifications.

CryoEM imaging, data processing and resolution assessment

To determine the atomic structure, low dose images of liquid-nitrogen-cooled, frozen hydrated BPP-1 were recorded on Kodak SO-163 film on an FEI Titan Krios cryo electron microscope operated at 300 kV with dose of ~25e⁻¹/Å² on specimen. The nominal magnification of images is 59,000×, which was previously calibrated to be 57,660× using tobacco mosaic virus as a standard. Imaging condition was optimized by using parallel illumination and by minimizing beam tilt with a Coma-free alignment procedure.

895 films were recorded and scanned using Nikon Coolscan9000 scanners under step-size of 6.35 μm, corresponding to 1.1 Å/pixel on the specimen scale. Under-defocus value of these films were determined using CTFFIND (*Mindell and Grigorieff, 2003*), and based on the cross correlation coefficients from CTFFIND3, 340 films with defocus values of -0.8~ -2.27 μm were selected for further processing. Particles were selected by an in-house automatic procedure and followed by visual screening to remove particles near edge of films, and total 43,156 particles with an image size of 880 × 880 were boxed. Data processing and 3D reconstruction were accomplished with an integrative approach using FREALIGN (*Grigorieff, 2007*) and eLite3D (*Zhang et al., 2010b*) on high performance computers including graphics processing units (GPU) as previously described (*Zhang et al., 2010a*). For global search, 2× binned images were used to save time. At the end of each cycle of image processing, the effective resolution was



Video 5. Non-covalent chainmail formed by MCPs on the BPP-1 capsid. For clarity, only the P domain and the E-loop of MCP subunit are shown. These domains from different capsomers join in a head to tail fashion to form concatenated rings. Ball-and-stick models are produced from our atomic model of BPP-1 and those in the same ring are rendered in the same color. Related to **Figure 6**. DOI: [10.7554/eLife.01299.020](https://doi.org/10.7554/eLife.01299.020)

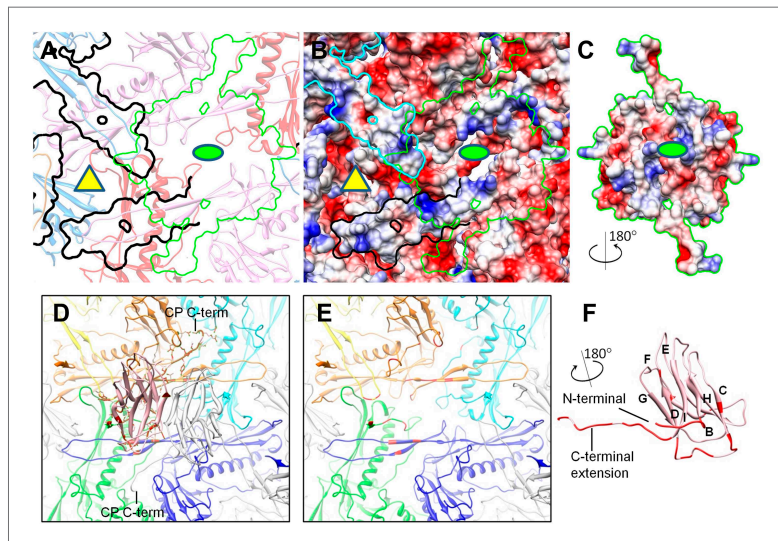


Figure 7. Interactions between CP and MCP. (A–C) The same views as **Figure 6C,D** are shown in (A) and (B), respectively, except for the addition of a green outline depicting the interaction interface between a CP dimer and its underlying MCPs. At the interaction interface with MCPs, the electrostatic potential surface of the CP dimer (C) is complementary to that of its underlying MCPs (B). (D–F) Details of the interactions (segments highlighted in red) between CP and MCP. (D) The two monomers in the CP dimer are shown as pink and grey ribbons and the five underlying MCPs as yellow, orange, cyan, blue, and green ribbons. Side chains involved in interactions with the pink CP are shown displayed as sticks. Note the extensive interactions between the CP C-terminal extension and three different MCP monomers. (E) The same as (D) but without the overlying CP dimer and sticks of interacting side chains to better reveal MCP segments involved in the interactions (red). (F) The bottom view of the pink CP monomer to better reveal its segments (red) involved in CP–MCP interactions, which include its N-terminal loop, C-terminal loop, and the loop connecting strands C and D in the jellyroll.

DOI: 10.7554/eLife.01299.021

determined and refinement for the next cycle would be carried out by including data up to this resolution. In the last refinement cycle, data up to a spatial frequency of $1/3.7 \text{ \AA}^{-1}$ was included for the refinement and 39,549 particles were used for the final reconstruction. To further improve signal/noise ratio, local averaging of seven CP or six hexameric MCP subunits (the pentameric MCP is slightly different from the hexameric MCP subunits and was not included in the averaging) in the asymmetric unit was performed as previously described (Zhang et al., 2010a).

The effective resolutions were estimated based on the 0.5 R-factors between the cryoEM density map [equivalent to $\text{FSC} \geq 0.143$ (Rosenthal and Henderson, 2003; Wolf et al., 2010) and the final atomic model calculated by Phenix (Adams et al., 2010). The calculated R-factors reached ~ 0.5 at 3.5 Å, 3.4 Å for the capsid density and the averaged CP and MCP densities, respectively (Table 1). These estimations are consistent with the structural features present in the maps (Figure 2, Videos 2–4). The capsid and averaged maps were filtered to $1/(3.4 \text{ \AA})$ and $1/(3.5 \text{ \AA})$ spatial frequency, respectively; and sharpened using a reverse B-factor of -200 \AA^2 (for capsid) or -250 \AA^2 (for averaged densities), which were estimated through a trial-and-error procedure by optimizing side-chain densities and noise level simultaneously. Visualization and segmentation of density maps were done with UCSF Chimera (Pettersen et al., 2004).

Atomic modeling and model refinement

Based on the averaged density maps of CP and MCP, we first built initial C_α and full atom models for CP and MCP with Coot (Emsley and Cowtan, 2004) without referring to any existing models

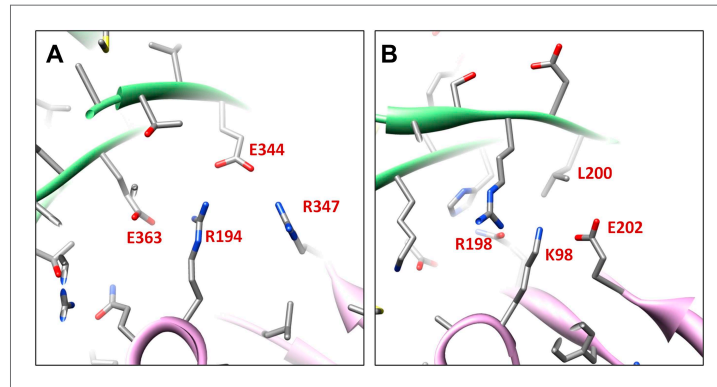


Figure 8. Salt bridges between adjacent HK97 gp5 molecules and lack of them in the corresponding positions in BPP-1 MCP molecules. (A) Close-up view of two adjacent HK97 gp5 subunits, showing salt bridges between Arg194, Arg347 and Glu363 and Glu344. (B) There are no such salt bridges in the corresponding regions in the BPP-1 MCP molecules.

DOI: [10.7554/eLife.01299.022](https://doi.org/10.7554/eLife.01299.022)

of other proteins. For MCP, the N- and C-termini were distinguished based on the 'Christmas tree' polarity of α helices and confirmed by landmark, bulky side chain densities. For CP, because there is no helix to be used to reveal the N- to C-terminal polarity, we determined the N-terminus using the side chain densities of some landmark amino acids, such as Phe27, Tyr 102, Tyr106, Tyr107 and Tyr133. The initial full atom models were regularized by constraining both Ramachandran geometry and secondary structures in Coot (Emsley and Cowtan, 2004) but without including hydrogen atoms.

These initial full atom models were iteratively refined using structural information of both amplitude and phase (from Fourier transform of the cryoEM maps) in the following three distinct (two automatic

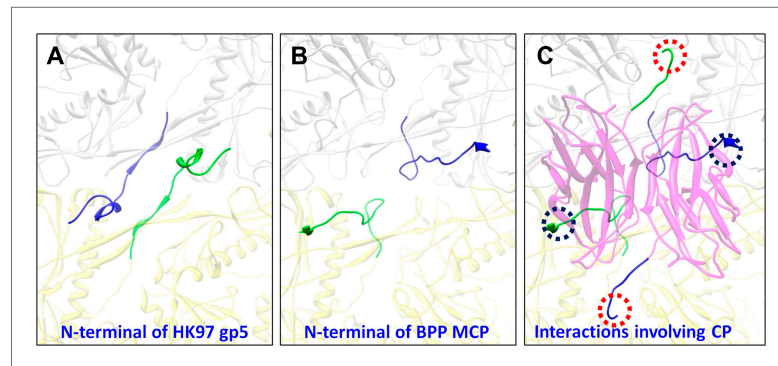


Figure 9. Interactions between CP and MCPs. (A) Ribbon models of two adjacent HK97 gp5 subunits, showing the interaction of the N-terminal loop (bright colored) with an adjacent gp5 (dimmed). (B) BPP-1 regions corresponding to that in (A) showing the lack of the HK97 type of interactions between the MCP N-terminal loop (bright colored) and adjacent MCP molecules (dimmed). (C) The same region in (B) but with the CP dimer (purple), showing that the N-terminal loop of MCP hydrogen-bonds to β -strand C of CP (black circles) to form an augmented 10-stranded β sheet, and that the C-terminal loop (red circles) of each CP has extensive interactions with nearby MCP molecules.

DOI: [10.7554/eLife.01299.023](https://doi.org/10.7554/eLife.01299.023)

and one manual) steps. The first step (automatic) was performed in Phenix (Adams *et al.*, 2010) using Ramachandran restraint. The second step (automatic) is to regularize the new model also using Phenix. To do so, hydrogens were added to all atoms of the model from the last refinement, and followed by regularization and removal of hydrogens. The latest models were refined iteratively until no further improvement was apparent based on both Ramachandran geometry and R-factors. Then in the third, manual step, amino acid residues with invalid Ramachandran backbone geometries were identified and their backbone Psi-Phi angles were manually corrected in Coot. This process of automatic- and manual model refinement steps was iterated until no further improvement on both Ramachandran geometry and R-factors was evident.

Atomic model of an asymmetric unit including seven MCPs and seven CPs was subsequently obtained by adjusting the refined MCP and CP models according to the density maps of different conformers. This model of the asymmetric unit was refined with Phenix under the constraints of Ramachandran geometry, secondary structures, and icosahedral symmetry (Adams *et al.*, 2010). Clashes at the molecular boundaries across different asymmetric units in the entire capsid were minimized by including icosahedral symmetry constraints in this refinement. The final atomic model of the full capsid was obtained after 14 cycles of refinement.

Because our BPP-1 MCP chain trace differs from that of the HK97 gp5 in the order of the α and β structural elements ('Results'), we made an extra cautious effort to verify our trace by swapping the α and β structural elements in our model to create an interchanged model that matches the HK97 topology. This interchanged model was then refined with Phenix (Adams *et al.*, 2010) for five cycles. Most of the side chains of the refined interchanged model do not match those in the cryoEM density, further confirming our de novo model (Figure 4—figure supplement 3).

Structure-based mutagenesis

To engineer BPP-1 MCP to match the topology of HK97 gp5, we swapped the order of the β - and α -elements in BPP-1 MCP (Figure 5—figure supplement 1C). In construct PM1, the β -element (peptide 169–241) was cut from wt *bbp17* gene and then pasted to the C-terminal end of α -element. Construct PM2 was obtained the same way except that the cut sites were shifted by three amino acid residues on both sides, resulting in a six residue-longer β -element (peptide 166–244). In both the constructs, the N-terminal end of the α -element was pasted to the C-terminal end of the N-element.

Plasmids expressing either wt *bbp17*, PM1 or PM2 genes *B. bronchiseptica* were transformed into RB50 cells transformed by electroporation. These transformed RB50 cells were grown on Bordet-Gengou agar containing 15% sheep blood, 25 μ g/ml streptomycin and 25 μ g/ml chloramphenicol. The protein expressing levels of the wt *bbp17*, PM1 and PM2 genes (each tagged with 6xhistidine at the C-terminus) in these RB50 cells were determined by Western blot. Single colonies were inoculated into Luria-Bertani media containing 25 μ g/ml streptomycin, 25 μ g/ml chloramphenicol and 10 mM nicotinic acid and grown at 37°C overnight. Cells in the amount of 1 ml at $OD_{600} = 1.0$ were pelleted and then grown in 2.5 ml of Stainer Scholte media with 25 μ g/ml streptomycin and 25 μ g/ml chloramphenicol to induce expression of the *bbp17*, PM1 or PM2 gene. Cells expressing each construct were grown for 3 and 6 hr. Equal amounts of the cells (by OD_{600}) were harvested and lysed by boiling in 1 \times SDS-PAGE loading buffer, and subsequently analyzed with SDS-PAGE. Western blot was done with a mouse monoclonal antibody against 6xhistidine as the primary antibody and a horse reddish peroxidase-conjugated goat anti-mouse antibody as the secondary antibody with an Amersham kit.

To obtain phage lysates for plaque assays, we generated a lysogen (BPP-1 Δ brt Δ bbp17) that has the *bbp17* gene deleted. Then, we transformed plasmids expressing either the wt *bbp17* gene, the PM1 or PM2 genes into the BPP-1 Δ brt Δ bbp17 lysogens. These lysogen cells were first grown at 37°C for 3 hr in Stainer Scholte medium containing 25 μ g/ml streptomycin and 25 μ g/ml chloramphenicol to induce to the Bvg⁺ phase, leading to the expression of the wt *bbp17* and the two mutants PM1 and PM2. Mitomycin C (0.2 μ g/ml) was then added to the cell cultures to induce phage production. After 3 hr, chloroform was added to the cultures, followed by vortexing and centrifugation to remove cellular debris. The resultant supernatants were collected for plaque assays on *B. bronchiseptica* RB50 cells transformed with plasmids expressing either wt *bbp17*, PM1 or PM2 gene. As above, these transformed RB50 cells were grown on Bordet-Gengou agar containing 15% sheep blood, 25 μ g/ml streptomycin and 25 μ g/ml chloramphenicol.

Visualization

CryoEM density maps, atomic models and surface charge properties were visualized with Chimera (Pettersen *et al.*, 2004).

Accession numbers

The cryoEM density map and the atomic model of BPP-1 have been deposited to databanks with accession numbers EMD (5764, 5765, 5766) and PDB (3J4U), respectively.

Acknowledgements

We thank Pavel Afonine and Paul Adams for advice on using Phenix to refine the atomic models. We acknowledge the use of the cryoEM facility in the Electron Imaging Center for Nanomachines by NIH (1S10RR23057 to ZHZ) and CNSI at UCLA.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	AI046420, GM071940	Z Hong Zhou
National Institutes of Health	AI069838	Jeff F Miller

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

XZ, LJ, EC, AH, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; HG, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article, Contributed unpublished essential data or reagents; WHH, AWN, Acquisition of data, Drafting or revising the article; JFM, ZHZ, Conception and design, Analysis and interpretation of data, Drafting or revising the article

Additional files

Major datasets

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Zhang X, Guo H, Jin L, Czornyj E, Hodes A, Hui WH, Nieh AW, Miller JF, Zhou ZH	2013	A new topology of the HK97-like fold revealed in <i>Bordetella</i> bacteriophage: non-covalent chainmail secured by jellyrolls	3J4U; http://www.pdb.org/pdb/search/structidSearch.do?structureId=3J4U	Publicly available at the Protein Data Bank (http://www.rcsb.org/pdb/).
Zhang X, Guo H, Jin L, Czornyj E, Hodes A, Hui WH, Nieh AW, Miller JF, Zhou ZH	2013	A new topology of the HK97-like fold revealed in <i>Bordetella</i> bacteriophage by cryoEM at 3.5A resolution	EMD-5764; http://www.ebi.ac.uk/pdbe/entry/EMD-5764	Publicly available at the Electron Microscopy Data Bank (http://http://www.ebi.ac.uk/pdbe/emdb/).
Zhang X, Guo H, Jin L, Czornyj E, Hodes A, Hui WH, Nieh AW, Miller JF, Zhou ZH	2013	A new topology of the HK97-like fold revealed in <i>Bordetella</i> bacteriophage by cryoEM at 3.5A resolution	EMD-5765; http://www.ebi.ac.uk/pdbe/entry/EMD-5765	Publicly available at the Electron Microscopy Data Bank (http://http://www.ebi.ac.uk/pdbe/emdb/).
Zhang X, Guo H, Jin L, Czornyj E, Hodes A, Hui WH, Nieh AW, Miller JF, Zhou ZH	2013	A new topology of the HK97-like fold revealed in <i>Bordetella</i> bacteriophage by cryoEM at 3.5A resolution	EMD-5766; http://www.ebi.ac.uk/pdbe/entry/EMD-5766	Publicly available at the Electron Microscopy Data Bank (http://http://www.ebi.ac.uk/pdbe/emdb/).

References

- Abad-Zapatero C, Abdel-Meguid SS, Johnson JE, Leslie AGW, Rayment I, Rossmann MG, et al. 1980. Structure of southern bean mosaic virus at 2.8 Å resolution. *Nature* **286**:33–39. doi: [10.1038/286033a0](https://doi.org/10.1038/286033a0).
- Abrescia NG, Cockburn JJ, Grimes JM, Sutton GC, Diprose JM, Butcher SJ, et al. 2004. Insights into assembly from structural analysis of bacteriophage PRD1. *Nature* **432**:68–74. doi: [10.1038/nature03056](https://doi.org/10.1038/nature03056).
- Abrescia NG, Bamford DH, Grimes JM, Stuart DI. 2012. Structure unifies the viral universe. *Annu Rev Biochem* **81**:795–822. doi: [10.1146/annurev-biochem-060910-095130](https://doi.org/10.1146/annurev-biochem-060910-095130).
- Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, et al. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst D* **66**:213–21. doi: [10.1107/S0907444909052925](https://doi.org/10.1107/S0907444909052925).
- Akita F, Chong KT, Tanaka H, Yamashita E, Miyazaki N, Nakaishi Y, et al. 2007. The crystal structure of a virus-like particle from the hyperthermophilic archaeon *Pyrococcus furiosus* provides insight into the evolution of viruses. *J Mol Biol* **368**:1469–83. doi: [10.1016/j.jmb.2007.02.075](https://doi.org/10.1016/j.jmb.2007.02.075).
- Baker ML, Hryc CF, Zhang Q, Wu W, Jakana J, Haase-Pettingell C, et al. 2013. Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling. *Proc Natl Acad Sci USA* **110**:12301–6. doi: [10.1073/pnas.1309947110](https://doi.org/10.1073/pnas.1309947110).
- Bamford DH, Grimes JM, Stuart DI. 2005. What does structure tell us about virus evolution? *Curr Opin Struct Biol* **15**:655–63. doi: [10.1016/j.sbi.2005.10.012](https://doi.org/10.1016/j.sbi.2005.10.012).
- Chelvanayagam G, Heringa J, Argos P. 1992. Anatomy and evolution of proteins displaying the viral capsid jellyroll topology. *J Mol Biol* **228**:220–42. doi: [10.1016/0022-2836\(92\)90502-B](https://doi.org/10.1016/0022-2836(92)90502-B).
- 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**:4347–52. doi: [10.1073/pnas.0915008107](https://doi.org/10.1073/pnas.0915008107).
- Duda RL. 1998. Protein chainmail: catenated protein in viral capsids. *Cell* **94**:55–60. doi: [10.1016/S0092-8674\(00\)81221-0](https://doi.org/10.1016/S0092-8674(00)81221-0).
- Emsley P, Cowtan K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**:2126–32. doi: [10.1107/S0907444904019158](https://doi.org/10.1107/S0907444904019158).
- Fokine A, Leiman PG, Shneider MM, Ahvazi B, Boeshans KM, Steven AC, et al. 2005. Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc Natl Acad Sci USA* **102**:7163–8. doi: [10.1073/pnas.0502164102](https://doi.org/10.1073/pnas.0502164102).
- Fuller DN, Raymer DM, Rickgauer JP, Robertson RM, Catalano CE, Anderson DL, et al. 2007. Measurements of single DNA molecule packaging dynamics in bacteriophage lambda reveal high forces, high motor processivity, and capsid transformations. *J Mol Biol* **373**:1113–22. doi: [10.1016/j.jmb.2007.09.011](https://doi.org/10.1016/j.jmb.2007.09.011).
- Gertsman I, Fu CY, Huang R, Komives EA, Johnson JE. 2010. Critical salt bridges guide capsid assembly, stability, and maturation behavior in bacteriophage HK97. *Mol Cell Proteomics* **9**:1752–63. doi: [10.1074/mcp.M000039-MCP201](https://doi.org/10.1074/mcp.M000039-MCP201).
- Grigorieff N. 2007. FREALIGN: high-resolution refinement of single particle structures. *J Struct Biol* **157**:117–25. doi: [10.1016/j.jsb.2006.05.004](https://doi.org/10.1016/j.jsb.2006.05.004).
- Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, Zientara S, et al. 1998. The atomic structure of the bluetongue virus core. *Nature* **395**:470–8. doi: [10.1038/26694](https://doi.org/10.1038/26694).
- Guo H, Tse LV, Barbalat R, Sivaamnuaihorn S, Xu M, Doulatov S, et al. 2008. Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Mol Cell* **31**:813–23. doi: [10.1016/j.molcel.2008.07.022](https://doi.org/10.1016/j.molcel.2008.07.022).
- Harrison SC, Olson AJ, Schutt CE, Winkler FK, Bricogne G. 1978. Tomato bushy stunt virus at 2.9 Å resolution. *Nature* **276**:368–373. doi: [10.1038/276368a0](https://doi.org/10.1038/276368a0).
- Hogle JM, Chow M, Filman DJ. 1985. Three-dimensional structure of poliovirus at 2.9 Å resolution. *Science* **229**:1358–65. doi: [10.1126/science.2994218](https://doi.org/10.1126/science.2994218).
- Jiang W, Baker ML, Jakana J, Weigele PR, King J, Chiu W. 2008. Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature* **451**:1130–4. doi: [10.1038/nature06665](https://doi.org/10.1038/nature06665).
- Krupovic M, Bamford DH. 2011. Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. *Curr Opin Virol* **1**:118–24. doi: [10.1016/j.coviro.2011.06.001](https://doi.org/10.1016/j.coviro.2011.06.001).
- Lander GC, Evilevitch A, Jeembaeva M, Potter CS, Carragher B, Johnson JE. 2008. Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM. *Structure* **16**:1399–406. doi: [10.1016/j.str.2008.05.016](https://doi.org/10.1016/j.str.2008.05.016).
- Liemann S, Chandran K, Baker TS, Nibert ML, Harrison SC. 2002. Structure of the reovirus membrane-penetration protein, Mu1, in a complex with its protector protein, Sigma3. *Cell* **108**:283–95. doi: [10.1016/S0092-8674\(02\)00612-8](https://doi.org/10.1016/S0092-8674(02)00612-8).
- Liu H, Jin L, Koh SB, Atanasov I, Schein S, Wu L, et al. 2010. Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science* **329**:1038–43. doi: [10.1126/science.1187433](https://doi.org/10.1126/science.1187433).
- Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, et al. 2002. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science* **295**:2091–4. doi: [10.1126/science.1067467](https://doi.org/10.1126/science.1067467).
- Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, Baker S, et al. 2004. Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol* **186**:1503–17. doi: [10.1128/JB.186.5.1503-1517.2004](https://doi.org/10.1128/JB.186.5.1503-1517.2004).
- Mathieu M, Petitpas I, Navaza J, Lepault J, Kohli E, Pothier P, et al. 2001. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. *EMBO J* **20**:1485–97. doi: [10.1093/emboj/20.7.1485](https://doi.org/10.1093/emboj/20.7.1485).

- McKenna R, Xia D, Willingmann P, Ilag LL, Krishnaswamy S, Rossmann MG, et al. 1992. Atomic structure of single-stranded DNA bacteriophage phi X174 and its functional implications. *Nature* **355**:137–43. doi: [10.1038/355137a0](https://doi.org/10.1038/355137a0).
- Mindell JA, Grigorieff N. 2003. Accurate determination of local defocus and specimen tilt in electron microscopy. *J Struct Biol* **142**:334–47. doi: [10.1016/S1047-8477\(03\)00069-8](https://doi.org/10.1016/S1047-8477(03)00069-8).
- Oksanen HM, Pietilä MK, Sencilo A, Atanasova NS, Roine E, Bamford DH. 2012. Virus universe: can it be constructed from a limited number of viral architectures. In: Witzany G, editor. *Viruses: essential agents of life*. Netherlands: Springer. p. 83–105. doi: [10.1007/978-94-007-4899-6_5](https://doi.org/10.1007/978-94-007-4899-6_5).
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* **25**:1605–12. doi: [10.1002/jcc.20084](https://doi.org/10.1002/jcc.20084).
- Rosenthal PB, Henderson R. 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol* **333**:721–45. doi: [10.1016/j.jmb.2003.07.013](https://doi.org/10.1016/j.jmb.2003.07.013).
- Rossmann MG, Arnold E, Erickson JW, Frankenberger EA, Griffith JP, Hecht HJ, et al. 1985. Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature* **317**:145–53. doi: [10.1038/317145a0](https://doi.org/10.1038/317145a0).
- Rossmann MG, Johnson JE. 1989. Icosahedral RNA virus structure. *Annu Rev Biochem* **58**:533–73. doi: [10.1146/annurev.bi.58.070189.002533](https://doi.org/10.1146/annurev.bi.58.070189.002533).
- Sternberg N, Weisberg R. 1977. Packaging of coliphage lambda DNA. II. The role of the gene D protein. *J Mol Biol* **117**:733–59. doi: [10.1016/0022-2836\(77\)90067-5](https://doi.org/10.1016/0022-2836(77)90067-5).
- Sutter M, Boehringer D, Gutmann S, Gunther S, Prangishvili D, Loessner MJ, et al. 2008. Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat Struct Mol Biol* **15**:939–47. doi: [10.1038/nsmb.1473](https://doi.org/10.1038/nsmb.1473).
- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**:4673–80. doi: [10.1093/nar/22.22.4673](https://doi.org/10.1093/nar/22.22.4673).
- Valegard K, Liljas L, Fridborg K, Unge T. 1990. The three-dimensional structure of the bacterial virus MS2. *Nature* **345**:36–41. doi: [10.1038/345036a0](https://doi.org/10.1038/345036a0).
- Vogel C, Morea V. 2006. Duplication, divergence and formation of novel protein topologies. *Bioessays* **28**:973–8. doi: [10.1002/bies.20474](https://doi.org/10.1002/bies.20474).
- Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE. 2000. Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289**:2129–33. doi: [10.1126/science.289.5487.2129](https://doi.org/10.1126/science.289.5487.2129).
- Wolf M, Garcea RL, Grigorieff N, Harrison SC. 2010. Subunit interactions in bovine papillomavirus. *Proc Natl Acad Sci USA* **107**:6298–303. doi: [10.1073/pnas.0914604107](https://doi.org/10.1073/pnas.0914604107).
- Zhang X, Jin L, Fang Q, Hui WH, Zhou ZH. 2010a. 3.3 Å cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell* **141**:472–82. doi: [10.1016/j.cell.2010.03.041](https://doi.org/10.1016/j.cell.2010.03.041).
- Zhang X, Zhang X, Zhou ZH. 2010b. Low cost, high performance GPU computing solution for atomic resolution cryoEM single-particle reconstruction. *J Struct Biol* **172**:400–6. doi: [10.1016/j.jsb.2010.05.006](https://doi.org/10.1016/j.jsb.2010.05.006).
- Zhang X, Guo H, Jin L, Miller J, Zhou H. 2012. Atomic structure of Bordetella Bacteriophage reveals a jellyroll fold in cement protein and a topologically distinct HK97-like fold in major capsid protein. *Microscopy and Microanalysis* **18**:72–73. doi: [10.1017/S1431927612002218](https://doi.org/10.1017/S1431927612002218).
- Zubieta C, Schoehn G, Chroboczek J, Cusack S. 2005. The structure of the human adenovirus 2 penton. *Mol Cell* **17**:121–35. doi: [10.1016/j.molcel.2004.11.041](https://doi.org/10.1016/j.molcel.2004.11.041).