ANNUAL REVIEWS

Annual Review of Microbiology

Accelerated Evolution by Diversity-Generating Retroelements

Benjamin R. Macadangdang,[1,2] Sara K. Makanani,[2,3,4] and Jeff F. Miller[2,4,5]

[1] Division of Neonatology and Developmental Biology, Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, California, USA; email: bmacadangdang@mednet.ucla.edu

[2] California NanoSystems Institute, University of California, Los Angeles, California, USA

[3] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California, USA; email: saramakanani@ucla.edu

[4] Department of Microbiology, Immunology and Molecular Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA; email: jfmiller@ucla.edu

[5] Molecular Biology Institute, University of California, Los Angeles, California, USA

Annu. Rev. Microbiol. 2022. 76:389–411

First published as a Review in Advance on June 1, 2022

The *Annual Review of Microbiology* is online at micro.annualreviews.org

https://doi.org/10.1146/annurev-micro-030322-040423

Keywords

diversity-generating retroelements, evolution, mutagenesis, retroelements, adaptation, genetic variation

Abstract

Diversity-generating retroelements (DGRs) create vast amounts of targeted, functional diversity by facilitating the rapid evolution of ligand-binding protein domains. Thousands of DGRs have been identified in bacteria, archaea, and their respective viruses. They are broadly distributed throughout the microbial world, with enrichment observed in certain taxa and environments. The diversification machinery works through a novel mechanism termed mutagenic retrohoming, whereby nucleotide sequence information is copied from an invariant DNA template repeat (TR) into an RNA intermediate, selectively mutagenized at TR adenines during cDNA synthesis by a DGR-encoded reverse transcriptase, and transferred to a variable repeat (VR) region within a variable-protein gene (54). This unidirectional flow of information leaves TR-DNA sequences unmodified, allowing for repeated rounds of mutagenic retrohoming to optimize variable-protein function. DGR target genes are often modular and can encode one or more of a wide variety of discrete functional domains appended to a diversifiable ligand-binding motif. Bacterial variable proteins often localize to cell surfaces, although a subset appear to be cytoplasmic, while phage-encoded DGRs commonly diversify tail fiber–associated receptor-binding proteins. Here, we provide a comprehensive review of the mechanism and consequences of accelerated protein evolution by these unique and beneficial genetic elements.

Contents

INTRODUCTION

Genetic variation arises within populations over successive generations through distinct processes that include mutation, recombination, migration of alleles, and drift (32). Natural selection acts on this existing genetic variation to modulate diversity (70). Mutational effects on fitness range from beneficial to deleterious; however, most mutations appear to be either neutral or harmful (57). Efficient maintenance of genomic stability (i.e., preservation of genomic material) can enhance reproductive success, particularly in stable environments for which an organism is well-suited. Conversely, under nonoptimal conditions commonly observed in dynamic environments, mutations can promote adaptation. Therein lies the double-edged nature of mutagenesis: Beneficial variability must be tightly balanced with the potential for loss of fitness (26).

Mutations are generally assumed to arise independently of their potential fitness effects (91), and therefore, mutation rates can largely influence the timescale of adaptation (26). To this end, numerous strategies have evolved to promote rapid and targeted functional diversification. In vertebrates, these diversification mechanisms range from large-scale rearrangements such as chromatin diminution (71) and genome remodeling (43) to site-specific mechanisms such as V(D)J recombination (19) and somatic hypermutation in vertebrates (79). In bacteria these strategies include integron-mediated recombination (65) and DNA inversions that mediate phase variation

(94). Creating hot spots maximizes the likelihood of acquiring mutations that confer a selective advantage while minimizing the potential costs of hypermutation.

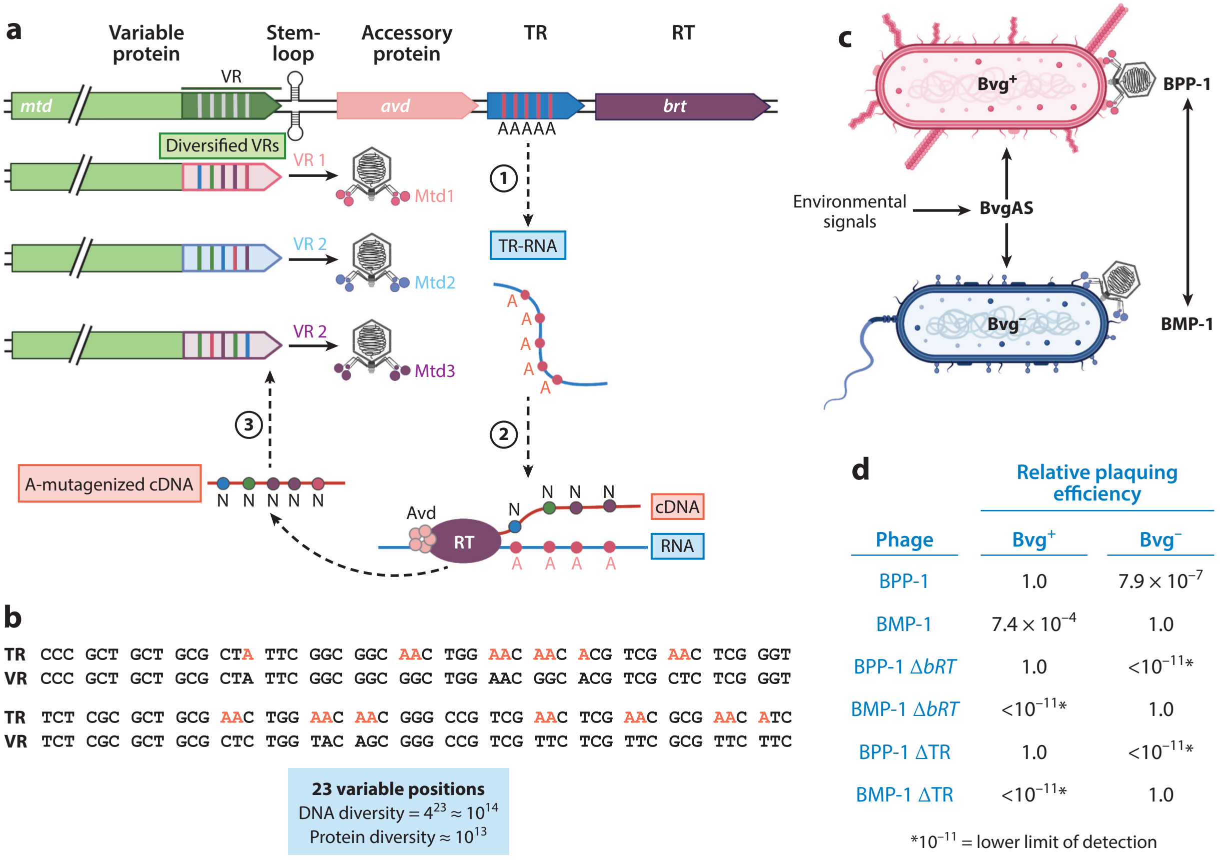
As one of the more vivid examples of targeted functional diversification, diversity-generating retroelements (DGRs) are capable of accelerating evolution by rapidly mutagenizing protein-encoding target genes (25, 54). DGRs were originally discovered in a *Bordetella* phage (54), and nearly 60,000 of them have now been uncovered in bacteria, archaea, and their viruses (81, 86, 97). In a subset of organisms longitudinally sampled in their natural environments, DGRs were responsible for more than 10% of all genomic amino acid changes (86). DGRs operate through a unique, unidirectional, copy-and-replace mechanism termed mutagenic retrohoming, which requires at least three components to operate: a target gene containing a variable repeat (VR) that is the recipient of mutagenesis, an invariant (usually intergenic) template repeat (TR) that is nearly identical to the VR, and a DGR-encoded reverse transcriptase (RT) (**Figure 1*a***). Protein diversity is created through a site-specific mechanism in which VR residues that correspond to adenine-containing codons in TR are selectively mutagenized (54) (**Figure 1*b***). During this diversification process, a TR-derived RNA intermediate serves as a template for cDNA synthesis by an error-prone, DGR-encoded RT in which adenine residues are disproportionately mismatched, resulting in random incorporation of any of the four dNTPs into cDNA at those positions (34) (**Figure 1*a*,*b***). The resulting adenine-mutagenized cDNA then replaces the VR sequence within the target gene, producing a variable protein that is diversified at specified sites in a specific domain. Because the TR is unaltered in this process, the VR can undergo repeated rounds of retrohoming and selection to evolve protein function.

The magnitude of diversity a DGR can generate is determined by the number and position of adenines in the TR. For example, a DGR found in *Elizabethkingia anopheles* harbors a TR with 56 adenines, providing the capacity to produce a VR with 4^{56} ($\sim 10^{33}$) unique DNA sequences, which corresponds to $\sim 10^{30}$ unique protein sequences (97). This diversification capability is many orders of magnitude greater than the number of unique immunoglobulins that can be produced by the vertebrate adaptive immune system ($\sim 10^{16}$, not including somatic hypermutation) (72). Of all the natural diversification systems known, DGRs have among the highest mutagenic capability; and remarkably, the machinery responsible for this staggering capacity is contained within a genetic footprint that usually spans about 3 kb. In this review, we delve deeper into these extraordinary retroelements, including their widespread prevalence in the microbial world and recent developments in our understanding of their unique diversification mechanism.

ECOLOGY AND EVOLUTION OF DGRs

Tropism Switching in *Bordetella* Bacteriophage

Bordetella pertussis and related species alternate between virulent (Bvg$^+$) and nonvirulent (Bvg$^-$) phases in response to environmental signals transduced through the BvgAS phosphorelay (9, 67) (**Figure 1*c***). Uniquely in the Bvg$^+$ state, *Bordetella* express adhesins, toxins, and other factors that facilitate colonization of the respiratory tract, including an autotransporter protein called pertactin (59). While investigating ligand-receptor interactions between pertactin and a *Bordetella* phage that preferentially infects Bvg$^+$ phase cells, called BPP-1 (Bvg-plus phage 1), Liu and colleagues (54) made a seminal observation: While the majority of BPP-1 progeny infected Bvg$^+$ cells as expected, a small number acquired the ability to efficiently infect Bvg$^-$ phase cells (termed BMP, or Bvg-minus phage), indicating that a tropism switch had occurred. They consistently detected these phages at a much higher frequency than would be expected if they arose through spontaneous mutation alone, suggesting there was an underlying mechanism at play (**Figure 1*d***). It was soon realized that underpinning this tropism-switching capability was a unique genetic element, a diversity-generating retroelement, that operated through mutagenic retrohoming to diversify



```
TR  CCC GCT GCT GCG CTA TTC GGC GGC AAC TGG AAC AAC ACG TCG AAC TCG GGT
VR  CCC GCT GCT GCG CTA TTC GGC GGC GGC TGG AAC GGC ACG TCG CTC TCG GGT

TR  TCT CGC GCT GCG AAC TGG AAC AAC GGG CCG TCG AAC TCG AAC GCG AAC ATC
VR  TCT CGC GCT GCG CTC TGG TAC AGC GGG CCG TCG TTC TCG TTC GCG TTC TTC
```

23 variable positions
DNA diversity = $4^{23} \approx 10^{14}$
Protein diversity ≈ 10^{13}

	Relative plaquing efficiency	
Phage	Bvg^{+}	Bvg^{-}
BPP-1	1.0	7.9×10^{-7}
BMP-1	7.4×10^{-4}	1.0
BPP-1 Δ*bRT*	1.0	$<10^{-11}$*
BMP-1 Δ*bRT*	$<10^{-11}$*	1.0
BPP-1 ΔTR	1.0	$<10^{-11}$*
BMP-1 ΔTR	$<10^{-11}$*	1.0

*10^{-11} = lower limit of detection

Figure 1

Overview of the BPP-1 DGR. (*a*) Components of the bacteriophage BPP-1 DGR and an outline of mutagenic retrohoming. BPP-1 DGR loci include *mtd*, which encodes the diversified receptor-binding protein that determines ligand specificity for infection; *avd*, which encodes a small, basic protein that forms a pentamer required for processive cDNA synthesis; the TR, which is partly homologous to VR sequences in *mtd* and is included in the TR-RNA intermediate used for cDNA synthesis; and *brt*, which encodes the BPP-1 DGR RT. Steps in mutagenic retrohoming include (①) production of the TR-RNA intermediate; (②) binding of TR-RNA to the Avd-RT complex to initiate cDNA synthesis via the error-prone DGR RT, which produces cDNA copies that are selectively mutagenized at adenines; and (③) integration and replacement of the parental VR allele with newly synthesized adenine-mutagenized cDNA. Gray bars in the VR (*top*) indicate sites corresponding to TR adenines that are subject to mutagenesis. (*b*) Comparison of BPP-1 VR and TR nucleotide sequences. Differences between VR and TR sequences occur exclusively at TR adenines (*red*). (*c*) DGR-mediated tropism switching by *Bordetella* phage BPP-1. In response to environmental signals, the *Bordetella* BvgAS phosphorelay mediates a transition between the Bvg^{+} (virulent) (*pink*) and Bvg^{-} (avirulent) (*blue*) phases, which is accompanied by major changes in expression of secreted and cell surface factors (67). Phage variants are shown with BPP (*pink*) or BMP (*blue*) tail fiber–associated Mtd molecules, indicating tropism for ligands specific to the Bvg^{+} or Bvg^{-} phase, respectively. (*d*) Phage tropism switching. *Bordetella* lysogens carrying integrated BPP-1 or BMP-1 prophage genomes, or mutant derivatives with deletions in *brt* (Δ*brt*) or TR (ΔTR), were induced with mitomycin C. After a single round of replication of the induced phage populations on their original hosts (Bvg^{+} phase *Bordetella* for BPP-1, Bvg^{-} phase *Bordetella* for BMP-1), relative plaquing efficiencies were determined on Bvg^{+} or Bvg^{-} phase *Bordetella* (54). Tropism switching requires both mutagenic retrohoming and the ability to generate Mtd variants capable of recognizing novel ligands in a functional manner. Since diversification is relatively rare following prophage induction, the majority of DGR-diversified genomes are packaged in phage particles with the parental Mtd specificity. Reinfection of the parental strain is required to detect phages that contain diversified genomes and the functional Mtd variants they encode. Abbreviations: Avd, accessory variability determinant; BMP, Bvg-minus phage; BPP, Bvg-plus phage; DGR, diversity-generating retroelement; Mtd, major tropism determinant; RT, reverse transcriptase; TR, template repeat; VR, variable repeat. Figure adapted from images created with BioRender.com.

the phage receptor-binding protein, Mtd (major tropism determinant), and accelerate the evolution of its ligand-binding domain (25, 54). Many observations made using the prototypical BPP-1 DGR as a model system have formed the foundation for core concepts that are now thought to be intrinsic to DGRs in general.

DGR Reverse Transcriptases and the Global Distribution of DGRs

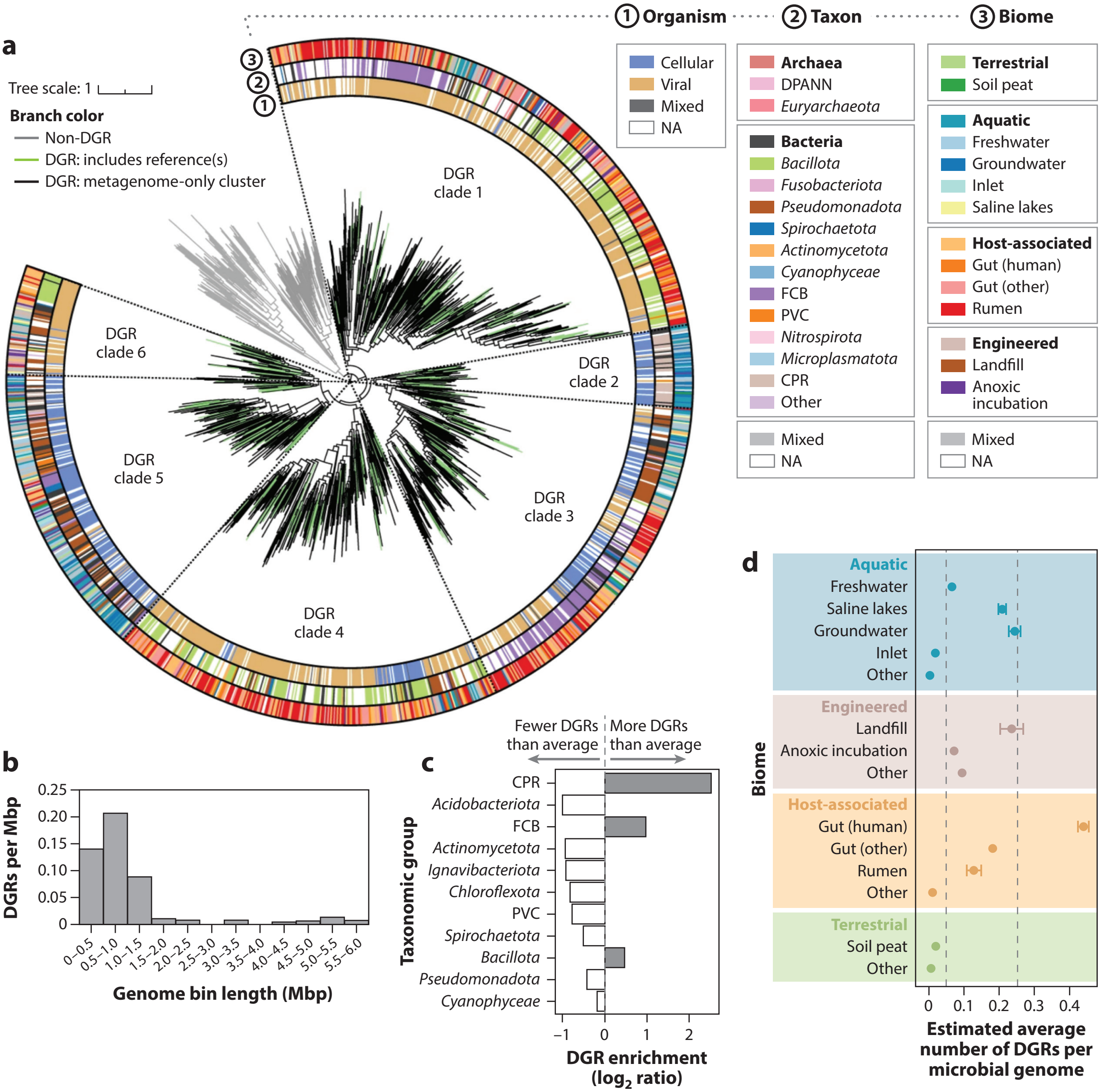
DGR RTs are most closely related to RTs encoded by group II introns and other retrons (82, 93), but they differ structurally due to some unique features. On average, DGR RTs are approximately 350 amino acids in length (interquartile range 327–394 amino acids) (86) and exhibit a conserved domain organization consisting of the palm and finger domains formed by RT motifs 1–7 (97). Like the RTs found in group II introns and non–long terminal repeat (non-LTR) retrotransposons, DGR RTs contain an extra motif, 2a (60), but they differ from those RTs in that they lack an upstream motif 0 (11). Notably, DGR RTs are devoid of RNase H domains that are typically present in other RTs and that hydrolyze RNA present in DNA/RNA hybrids (88, 97). As a result of their unique structure, DGR RTs cluster into a distinct, well-demarcated phylogenetic lineage that separates them from other RTs (25, 81, 86) (**Figure 2*a***). This also forms the basis for identifying new DGR RTs from genomic or metagenomic datasets using phylogenetic clustering (25), BLAST (basic local alignment search tool) (81), or hidden Markov model profiles (86). DNA sequences surrounding DGR RT hits are then scanned for near-identical repeats corresponding to VR/TR pairs. Using this strategy, an early study by Doulatov et al. (25) identified a subgroup of RTs in sequenced bacterial genomes that phylogenetically clustered on the same branch as the BPP-1 RT and were encoded adjacent to VR/TR pairs that differed at positions corresponding to TR adenines. From these analyses, the first homologous, non-phage-associated DGRs were discovered in genomes of distantly related bacteria, including *Bifidobacterium longum*, *Bacteroides thetaiotaomicron*, *Treponema denticola*, and several cyanobacterial species, in addition to bacteriophage genomes.

Having been discovered in bacteria and bacteriophages, DGRs initially appeared to be confined to the bacterial domain of life (7, 52). A subsequent sequencing study of subsurface sediments from a methane seep in the Santa Monica Basin, however, identified the first intact DGR in a virus predicted to infect archaea, and in two uncultivated nanoarchaea that each appeared to harbor multiple DGRs (80). Expanding on this effort, Paul et al. (81) analyzed a large metagenomic dataset derived from groundwater organisms isolated from an aquifer on the banks of the Colorado River, using serial filtration to collect cells of progressively smaller sizes (6, 13, 18). Previous reconstruction of the sampled genomes revealed that many of these small organisms belonged to the Candidate Phyla Radiation (CPR), a recently discovered branch that now accounts for 15% of the known diversity of the bacterial domain (13, 41), and to the archaeal superphylum consisting of "*Candidatus* Diapherotrites," "*Candidatus* Parvarchaeota," "*Candidatus* Aenigmarchaeota," "*Candidatus* Nanoarchaeota," and "*Candidatus* Nanohaloarchaeota" (DPANN) (80, 81). Along with extremely small cell sizes, many of these organisms have correspondingly small genomes and are missing genes involved in vital biosynthesis pathways, suggesting a requirement to form symbiotic relationships with larger hosts (31, 75). Within the CPR and DPANN groups, the presence of DGRs was highly correlated with smaller genome sizes (81) (**Figure 2*b***), prompting speculation that DGR-mediated diversification could be particularly advantageous for organisms in which selective pressures favor reduced genome size.

A remarkable observation following the discovery of DGRs in organisms that span vast phylogenetic distances, and two of the three domains of life, is that VRs differed from TRs almost exclusively at positions corresponding to TR adenines, suggesting that the mechanism of adenine-specific mutagenesis is highly conserved. Is this due to DGRs having originated from a single, ancient origin, or has adenine mutagenesis convergently evolved? With nearly 60,000 examples now reported (7, 10, 36, 69, 76, 80–82, 86, 87, 89, 97–99), all known DGR RTs form a monophyletic clade separate from other RTs (80, 81, 86) and have a branching pattern that mirrors that of the bacterial and archaeal domains (80, 81, 97). The accumulated evidence, therefore, points toward a single, ancient evolutionary origin of DGRs.

Ecology of DGRs

In one of the largest surveys of DGRs published, Roux et al. (86) analyzed 81,404 genomes and 9,467 metagenomes from samples collected across 90 ecological niches, identifying 32,321 DGRs



(*Caption appears on following page*)

Figure 2 (*Figure appears on preceding page*)

Distribution of DGRs across diverse taxa and biomes. (*a*) Phylogenetic analysis of RTs: 32,321 DGR RTs identified from a diverse dataset of genomic and metagenomic sequences were clustered, and a representative RT from each cluster along with a collection of non-DGR RTs were used to build a phylogenetic tree (86). Each DGR RT cluster was classified by (①) organism type (viral, cellular), (②) taxon, and (③) biome. (*b*) Distribution of DGRs across organisms with small genomes (81). Bacterial and archaeal genomes were reconstructed from organisms isolated by serial filtration of groundwater samples to collect progressively smaller cells. This dataset was enriched for CPR and DPANN organisms. DGR prevalence (number of DGRs per megabase) was calculated for each genome and binned according to genome size. (*c*) Enrichment of DGRs across taxonomic groups. A list of single-copy marker genes was used to estimate the number of genomes belonging to different taxa from the same collection of metagenomic assemblies as in panel *a*. The average frequency of DGRs was calculated for each taxonomic group, compared to the overall frequency of DGRs across the whole dataset, and reported as $\log_2$ enrichment ratios. All enrichments were statistically significant (chi-square test of independence corrected p value $< 10^{-10}$), except for the case of *Cyanophyceae*. (*d*) Average number of DGRs across biomes. Using the same collection of metagenomic assemblies as in panel *a*, a linear regression was calculated for each biome between the number of genomes and the number of DGRs in each metagenome assembly. The regression slope was then used to estimate the average number of DGRs present per genome, with error bars corresponding to the standard error of the slope estimation. Each biome type is represented by a color: aquatic, blue; engineered, brown; host-associated, orange; terrestrial, green. Abbreviations: CPR, Candidate Phyla Radiation; DGR, diversity-generating retroelements; DPANN, "*Candidatus* Diapherotrites," "*Candidatus* Parvarchaeota," "*Candidatus* Aenigmarchaeota," "*Candidatus* Nanoarchaeota," and "*Candidatus* Nanohaloarchaeota"; FCB, *Flavobacteria* (*Cytophagia*), *Fibrobacteres* (*Fibrobacterota*), *Chlorobi* (*Chlorobiota*), and *Bacteroidota*; NA, not assigned; PVC, *Planctomycetes* (*Planctomycetota*), *Verrucomicrobia*, and *Chlamydiae*; RT, reverse transcriptase. Panels *a*, *c*, and *d* adapted from Reference 86. Panel *b* adapted from Reference 81.

belonging to more than 1,500 bacterial and archaeal genera. In addition to CPR bacteria, phyla that appear to be particularly enriched in these retroelements include *Firmicutes* and the *Flavobacteria* (*Cytophagia*), *Fibrobacteres* (*Fibrobacterota*), *Chlorobi* (*Chlorobiota*), and *Bacteroidota* (FCB) supergroup (**Figure 2*c***), which codominate the intestinal microbiome (77). In an evaluation of the population diversity of DGR loci across taxa and ecosystems, single-nucleotide variants were identified in a majority of VR sequences, with 50–75% of DGRs showing clear signs of active diversification (86). The apparent lack of activity by the remaining elements is also interesting. It could be due to purifying selection and the rapid loss of new variants; the presence of host- or retroelement-encoded systems that control DGR activity; or mutations that disable the diversification machinery, leading to fixation of a particularly successful variable protein that provides an advantage in a stable niche.

DGRs have been identified in microbes that occupy terrestrial, subterrestrial, marine, freshwater, engineered, and host-associated environments (81, 86, 97). Besides the human gut, biomes consisting of groundwater reservoirs, saline lakes, and landfills were found to be enriched in organisms containing DGRs, compared to samples from other ecological niches such as terrestrial freshwater and soil (**Figure 2*d***). Using phylogenetic logistic regression (42), where a binary outcome (presence or absence of a DGR) can be predicted from ecological data taking into account the strength of phylogenetic relationships, Roux and colleagues (86) found that both phylogeny and ecology drive the distribution of DGRs observed in their dataset.

Horizontal Transfer of DGRs

With the recent establishment of large viral genome databases (16, 74), the distribution of phage-associated DGRs is coming into focus. More than half of all known DGRs are predicted to be of viral origin, either as free phage or prophage (10, 74, 86). The vast majority of these belong to the *Caudovirales* order, which are tailed, double-stranded DNA (dsDNA) bacteriophages, while no DGRs have been uncovered in single-stranded DNA (ssDNA) or RNA viruses (3, 61). Although *Caudovirales* inhabit a range of environmental niches, they are particularly abundant in metagenomic samples from the human gut (74), raising the possibility that sampling bias contributes to the current picture. Nonetheless, a prominent family of *Caudovirales*, namely the contractile tailed *Myoviridae*, includes a disproportionately large percentage of genomes (~84%) with DGRs, while other families contain few to none (74). Several features of the *Caudovirales* may make them particularly suitable hosts for DGRs. For instance, *Caudovirales* carry their genetic information on dsDNA and rely on high-fidelity polymerases to replicate their genomes. In contrast, faithful replication of RNA viruses would likely be hindered by an error-prone DGR RT. Additionally, members of the *Caudovirales* order typically have genome sizes that range between 45 and 155 kb, allowing them to accommodate accessory elements such as a 3-kb DGR (3). ssDNA viruses, on the other hand, limit their genomes to 1–12 kb. This may decrease susceptibility to spontaneous nicks, which can destroy ssDNA viral genomes (17, 27), but it leaves little room to encode additional fitness factors.

On the basis of their diversified target genes, phage DGRs generally fall into two groups. The majority appear to diversify tail fiber proteins involved in host recognition, as was shown for the *Bordetella* phage variable protein, Mtd (86) (see the section titled Variable Protein Structure and Function). Mutagenic retrohoming provides an efficient means for expanding the ligand specificity of a viral infection apparatus, promoting adaptation to dynamic host cells and the possibility for niche expansion. In contrast, a smaller subset of viral DGRs appear to diversify genes to the potential benefit of their bacterial or archaeal hosts (86). This places them within a class of viral accessory loci that are sometimes referred to as morons (because they add "more on" the phage genome) (45, 92). By definition, these genes are not primarily involved in phage functions but instead confer phenotypes such as antimicrobial resistance, toxin production, adhesin expression, or other traits that increase host fitness. Although the functional roles of moron DGRs have yet to be explored, their selective advantage is made possible by the ability to form lysogens, with prophage genomes maintained as chromosomally integrated elements or episomes that replicate in concert with host cells (12, 15, 21).

Not surprisingly, DGRs are also encoded by transposons, plasmids, integrons, and integrative and conjugative elements (ICEs) (28, 90). The best-characterized so far is a DGR carried on a 64-kb ICE in *Legionella pneumophila* strain Corby, a serogroup 1 clinical isolate from a patient with legionellosis (7, 51). With an estimated repertoire of 10^{19} unique polypeptides, the DGR diversifies a variable lipoprotein anchored in the outer leaflet of the outer membrane, with its C-terminal variable region surface exposed. Related DGRs were identified in other *L. pneumophila* clinical isolates, suggesting general utility for the surface display of such rapidly evolvable ligand-binding capabilities. On a global level, the distribution pattern of DGRs in nature is partly explained by the horizontal transfer of DGRs between organisms, both phylogenetically related and distinct, with selective retention in new hosts that can support and benefit from mutagenic retrohoming.

MUTAGENIC RETROHOMING

DGR-mediated diversification of target genes employs a mechanism of mutagenic retrohoming with features that are, in some cases, unlike anything else described. Our understanding of this process derives primarily from observations with *Bordetella* phage BPP-1, which has provided a tractable platform for mechanistic studies that apply broadly to DGRs in bacterial, archaeal, and phage genomes.

TR-RNA Is the Template for Information Transfer

The central roles of the TR element and the DGR-encoded RT in mutagenic retrohoming were first identified in mutagenesis studies reported by Liu et al. (54). Precise deletion of BPP-1 TR resulted in phages that remained fully infectious toward their original hosts but had lost the ability to switch tropism (**Figure 1*d***). Deletion of the *rt* locus, or mutations in the RT catalytic center, resulted in an identical phenotype—full infectivity but a loss of VR variability. These results, and the realization that variant VR sequences always differed at positions corresponding to TR adenines, while TR sequences remained constant, prompted the hypothesis that information is transferred from TR to VR in an RT-dependent manner that is somehow coupled to adenine-specific mutagenesis.

Unidirectional transfer of sequence information was first demonstrated by incorporating nucleotide substitutions into TR that were designed to create silent mutations if transferred to VR (25, 54). Subsequent selection for phage tropism variants revealed diversified VR sequences with synonymous mutations derived from TR, flanked by characteristic patterns of adenine mutagenesis. The observation of an essential role of the BPP-1 RT in DGR activity immediately suggested the involvement of an RNA intermediate, most likely derived from TR (54). Indeed, when a self-splicing group I intron from bacteriophage T4 (*td* intron) was introduced into the BPP-1 TR sequence, only precisely spliced exons were transferred to VR, accompanied by adenine mutagenesis (34). Since group I intron splicing only occurs at the RNA level (40, 49), these results provided genetic proof that DGRs operate through a mutagenetic retrotransposition process that uses TR-RNA as the template for reverse transcription.

cDNA Synthesis Is Template-Primed

The essential priming step for RNA-templated cDNA synthesis can occur through multiple pathways (**Figure 3**). During target-primed reverse transcription (TPRT), used by mobile retroelements ranging from group II introns in bacteria to non-LTR long interspersed elements (LINEs) in humans, homing and mobility occur through endonuclease cleavage at the target site, which generates a 3′-OH primer for initiating cDNA synthesis (20, 58, 101, 102) (**Figure 3*b***). As DGR RTs are most closely related to RTs of group II introns (**Figure 2*a***), it was initially assumed that DGR retrotransposition was similarly target-primed (33, 34). However, unlike RTs encoded by retroelements that use TPRT, DGR RTs lack recognizable nuclease domains, as do other conserved DGR-encoded proteins. This suggested that DGR-mediated cDNA synthesis (*a*) recruits host factors for target site cleavage, (*b*) uses a novel DGR-encoded endonuclease activity that had escaped detection, or (*c*) is primed by an entirely different mechanism (**Figure 3*c***).

A simple yet key prediction of the target-priming model is that cDNA synthesis requires the presence of target sequences. To test this, Naorem et al. constructed a self-replicating plasmid with inducible expression of BPP-1 DGR loci encoding Avd (accessory variability determinant), an accessory protein described below; RT; and TR tagged with a self-splicing Gp1 intron (73). Surprisingly, after induction of plasmid expression in a *Bordetella* strain that otherwise lacks BPP-1 prophage sequences, including the VR target sequence, cDNA products derived from spliced TR-RNA templates were readily observed. Both Avd and a catalytically active RT were essential for the synthesis of cDNA, which was shown to contain adenine-specific mutations. This demonstrated that DGR-directed cDNA synthesis can occur through a mechanism that does not require target sequences for priming.

To determine how cDNA synthesis was initiated in the absence of VR target sequences, 5′ RACE (rapid amplification of cDNA ends) and nested PCR were used to extend and amplify DGR RT-dependent cDNA products, which were cloned and sequenced (73, 84). Intriguingly, this analysis revealed the presence of chimeric RNA-cDNA molecules in which TR-derived, adenine-mutagenized cDNA sequences were fused to RNA sequences from the 140-nucleotide spacer region (sp) located between the TR and *rt*, but in reverse orientation (**Figure 3*c***). This suggested that spacer sequences in TR-RNA molecules had folded back upon themselves to initiate cDNA synthesis. Subsequent mutagenesis studies identified the priming nucleotide to be

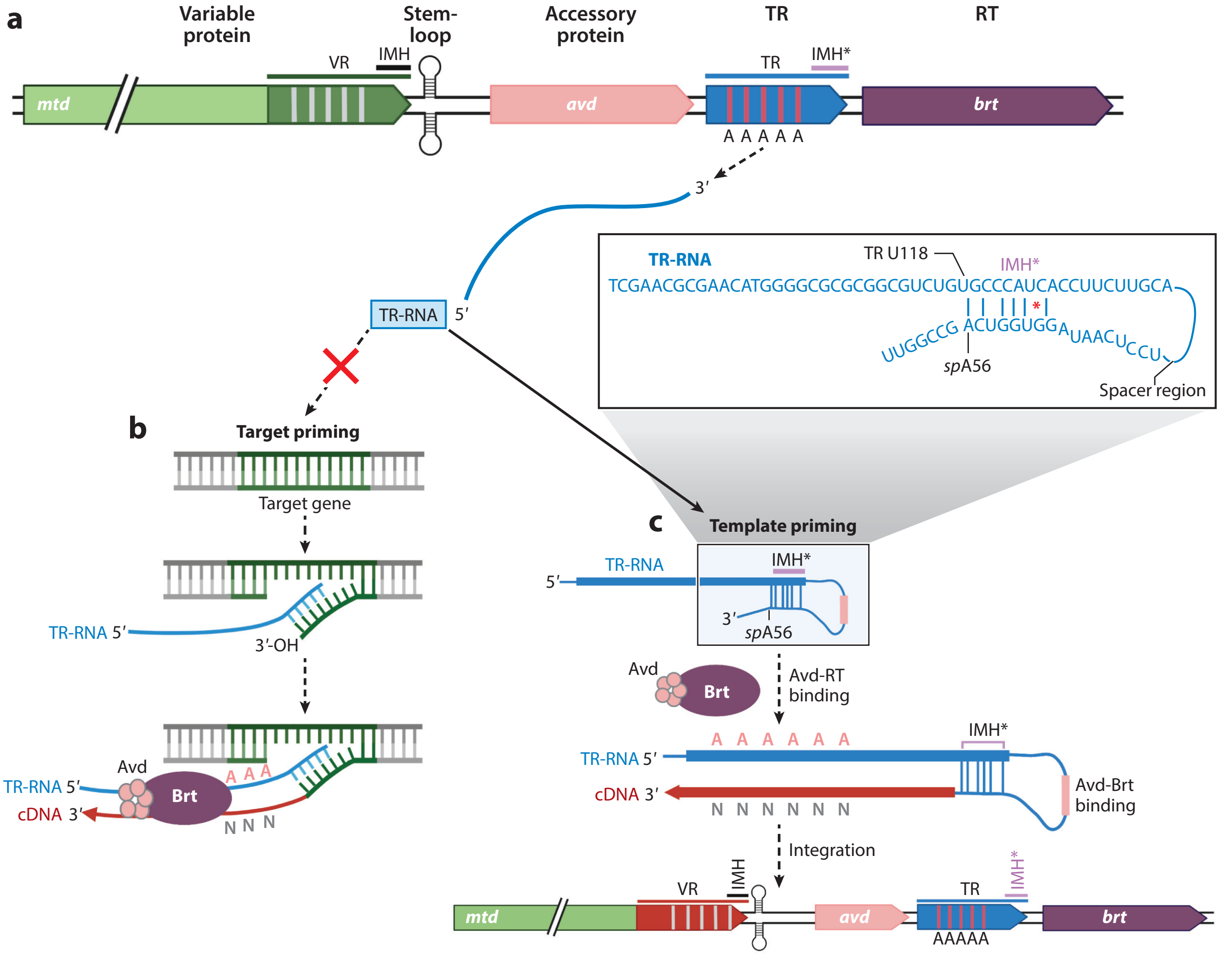


Figure 3

Mechanistic features of mutagenic retrohoming. (*a*) The DGR locus of phage BPP-1 is shown as in **Figure 1*a***, with the addition of IMH at the 3′ end of VR and the related, but not identical, IMH* sequence at the 3′ end of TR, which determine the directionality of information flow. DNA stem-loop structures flanking the 3′ end of *mtd* are also shown. (*b*) Target-primed reverse transcription. An endonuclease cleaves the antisense strand of the target gene, generating a 3′-OH for the initiation of cDNA synthesis. Hypothetically, the TR-RNA would then hybridize with target sequences and serve as a template for error-prone reverse transcription by the Avd-Brt polymerase complex. Experimental evidence showing that adenine-mutagenized cDNA synthesis occurs independently of the presence or absence of target sequences argues against a requirement for target priming during mutagenic retrohoming (73). (*c*) Template-primed reverse transcription. TR-RNA (*blue*) folds back upon itself and is stabilized by a region of complementarity between IMH* and sequences from the spacer region located between TR and *brt*. This confirmation positions *sp*A56 (spacer adenine 56 nucleotides downstream from TR) to prime cDNA synthesis using its available 2′-OH, or a 3′-OH released by removal of the 3′ TR-RNA tail. Avd-Brt binding to a recognition sequence in the TR-RNA spacer region is required for processive polymerization. The resulting adenine-mutagenized cDNA (*red*) integrates and replaces the VR parental allele through an unknown mechanism. Experimental evidence to date is consistent with template priming as the mechanism responsible for the initiation of cDNA synthesis during mutagenic retrohoming (37, 73). Abbreviations: Avd, accessory variability determinant; BPP-1, Bvg-plus phage 1; Brt, BPP-1 DGR RT; DGR, diversity-generating retroelements; IMH, initiation of mutagenic homing; *mtd*, major tropism determinant; RT, reverse transcriptase; TR, template repeat; VR, variable repeat. Figure adapted from images created with BioRender.com.

*sp*A56: an adenine residue located 56 nucleotides downstream from the end of VR/TR homology. A short region of internal complementarity in TR-RNA, between the IMH* sequence at the end of the TR and downstream spacer sequences, positions *sp*A56 for priming (73) (**Figure 3*c***). Two possibilities can be imagined for the priming reaction itself. Naorem et al. (73) presented evidence that the TR-RNA template was cleaved at *sp*A56, in what appeared to be an RT-dependent manner, to generate a 3′-OH to prime cDNA synthesis. An alternative possibility, supported by studies from an in vitro reconstituted system described below (37), is that cDNA synthesis is initiated at the 2′-OH of *sp*A56, resulting in a branched RNA-cDNA structure that may be subject to further processing. Although additional work is needed to resolve these and other details, the available evidence clearly supports a target-DNA-independent, template-RNA-primed mechanism for the initiation of cDNA synthesis by the BPP-1 DGR.

In Vitro Reconstitution of Mutagenic cDNA Synthesis

Adenine-specific mutagenesis of target DNA sequences is a unique and conserved hallmark of mutagenic retrohoming. To understand the biochemistry of this process, Handa et al. (37, 39) developed an in vitro system using purified, reconstituted BPP-1 DGR components that faithfully recapitulates DGR-mediated cDNA synthesis in vitro. This provides an outstanding opportunity to probe requirements for cDNA synthesis and the biochemical basis for selective infidelity during reverse transcription.

In addition to RT and TR-RNA, BPP-1 DGR activity requires a small basic protein called Avd, which is encoded directly upstream of TR (4, 34, 37, 54, 73) (**Figures 1*a*** and **3*a***). Avd forms a positively charged pentameric barrel that interacts with RT in a manner that is critical for function (4), and it is one of several accessory factors that are differentially conserved among subsets of DGRs (97). To reconstitute reverse transcription in vitro, Handa and colleagues (37) combined Avd-RT complexes coexpressed and purified from *Escherichia coli* with RNA transcripts containing the TR and adjacent sequences. In the presence of dNTPs, these components were necessary and sufficient to generate template-primed, covalently linked RNA-cDNA molecules that initiated cDNA synthesis at precisely the same nucleotide, *sp*A56, as occurs in vivo (73) (**Figure 3*c***). Just as remarkably, cDNAs were mutagenized at positions corresponding to TR adenines, with about 50% of adenines substituted per cDNA molecule, similar to what is observed in *Bordetella* (25, 73). This lends in vitro support for the template-primed model of DGR retrotransposition and shows that adenine-specific mutagenesis is an intrinsic property of the Avd-RT complex.

Avd Targets Reverse Transcription to TR-RNA

Deletion mutations in *avd* eliminate mutagenic retrohoming by the BPP-1 phage DGR (4, 25), and structure-guided point mutations that disrupt Avd-RT complex formation confer DGR-null phenotypes in vivo (4). The requirement of Avd for DGR activity is best understood in light of recent evidence that it is responsible for targeting mutagenic reverse transcription specifically to TR-RNA templates (37).

When presented with a heterologous (non-DGR) RNA transcript in vitro, Avd-RT is incapable of cDNA synthesis in the absence of a DNA primer. When a DNA primer is added, only short cDNA products are produced, indicating a lack of processivity of the polymerase complex. The mechanism for restricting processive, template-primed cDNA synthesis to TR-RNA became clear through a combination of RNase protection and mutagenesis studies showing that a 26-nucleotide sequence in the spacer region forms an Avd-binding site that targets the Avd-RT complex to TR-RNA and positions it to initiate cDNA synthesis at the correct position (37) (**Figure 3*c***). In addition to increasing the efficiency of self-priming, targeting DGR RT activity to the correct RNA template may also protect the host genome from damage. Reverse transcription of cellular mRNAs can lead to pseudogene formation following incorporation of cDNA products into host genomes, as documented for retrotransposons including LINEs and Ty elements (24, 30). Considering the mutagenic activity of DGR-directed reverse transcription, the consequences of such events could be particularly damaging to host genes.

DGR Reverse Transcriptases Are Remarkably Promiscuous

The extraordinary promiscuity of DGR RTs is orders of magnitude greater than that of any other characterized RNA- or DNA-dependent DNA polymerase. In relation to enzymes commonly used in the laboratory, the BPP-1 RT is about 1,000,000 times more error prone than Q5 polymerase (83). Compared to the error rate of HIV-1 RT, which is well known for promiscuity leading to escape from immunosurveillance and drug resistance, the error rate of DGR RT is approximately 10,000 times higher (2). Although error-prone DNA synthesis is well described, the exact mechanism of *selective infidelity* at adenines displayed by DGR RTs is perhaps the most intriguing unanswered question regarding their function.

Using single dNTP primer extension assays, the catalytic efficiency ($k_{\mathrm{cat}}/K_{\mathrm{m}}$) of the Avd-RT complex was measured in vitro (39). Like most polymerases, BPP-1 RT has a low catalytic efficiency for misincorporation, with the lowest efficiency observed at template adenines. Unlike high-fidelity enzymes, which have a greatly ($\sim10^5$-fold) improved efficiency for correct versus incorrect incorporation, Avd-RT has only a slightly higher catalytic efficiency for correct incorporation than it does for misincorporation, with the smallest difference (~20-fold) observed at template adenines. The adenine specificity of misincorporation by Avd-RT was probed in vitro using nucleobase analogs at either the N1, C2, or C6 positions of the purine ring, which differ between guanine, which is faithfully copied, and adenine, which promotes promiscuity. Of these positions, C6 had the greatest effect on fidelity. The presence of a carbonyl group, as occurs in guanine, significantly decreased misincorporation, but when an amine was present at C6, as in adenine, misincorporation occurred at a similar rate as when there is no side group at all. In mutagenesis studies aimed at identifying BPP-1 RT residues that modulate infidelity, Arg-74 and Ile-181 were found to promote adenine mutagenesis (39). These amino acids are predicted by in silico modeling to have counterparts in HIV RT that nonspecifically stabilize incoming dNTPs. Although the analysis is still at an early stage, the picture that emerges is that the unusually low catalytic efficiency of the BPP-1 DGR RT, in combination with structural features of the active site, is intimately tied to the ability to carry out adenine-selective mutagenesis. The availability of high-resolution atomic structures of the Avd-RT–TR-RNA riboprotein complex in the presence and absence of dNTPs will be essential for understanding the biophysical basis of this unique enzymatic behavior.

Mutagenic Retrohoming Is Reiterative

The V-J or V-D-J rearrangements that generate diversity in T cell receptor or antibody genes can produce repertoires of approximately 10^{16} unique polypeptides, of which a small subset experience antigen-driven selection, amplification, and further differentiation (72). Although the amount of diversity this system can generate is impressive, the required genetic rearrangements are irreversible, and consequently the mechanism can only operate once during the life span of a lymphocyte. In contrast, DGR-mediated diversification differs in several ways. The first involves efficiency—the genetic machinery required for mutagenic retrohoming is generally contained within ~3 kb, compared to the ~950,000 kb size of the human heavy chain locus alone (64). But despite their compact footprint, DGRs can generate far greater diversity, exceeding 10^{25}

polypeptide sequences in many cases (97). Second, and perhaps most importantly, mutagenic retrohoming is reiterative, and this allows protein function to be optimized by repeated rounds of diversification and selection.

The reiterative capacity of DGRs is ensured by two mechanistic features of mutagenic retrohoming (**Figure 3*a*,*c***): (*a*) the strict unidirectionality of information flow, from TR to VR, leaves the TR-DNA sequence unaltered, and (*b*) after successfully diversifying a VR, all essential *cis*-acting sites and *trans*-acting factors remain intact and available for subsequent rounds of retrohoming. The directionality of information flow is maintained by sequences located directly downstream from the VR and TR (25, 35, 73). These include the IMH (initiation of mutagenic homing) region, which in BPP-1 consists of a 21-bp sequence at the 3′ end of the VR that differs at five positions from its corresponding TR sequence, the IMH* (**Figure 3*a*,*c***). The difference between the IMH and IMH* is essential for directionality—mutating the VR IMH to resemble the IMH* extinguished retrohoming activity, while converting the TR IMH* sequence to IMH led to the diversification of adenines in TR (25). Immediately downstream from the BPP-1 IMH element is a 20-bp sequence that forms a hairpin or cruciform structure in DNA. Mutations that change the four-nucleotide loop or disrupt the 8-bp stem significantly decrease the efficiency of retrohoming, most likely by affecting the cDNA integration step that follows reverse transcription (35). Although mechanistic requirements for assimilating newly synthesized cDNA into a VR have yet to be defined, stem-loop structures are commonly observed to follow VR sequences in diverse DGRs (97). Importantly, sequencing VRs that have undergone recent diversification events clearly shows that IMH, IMH*, and stem-loop structures are precisely regenerated after every round of mutagenic retrohoming (25, 54).

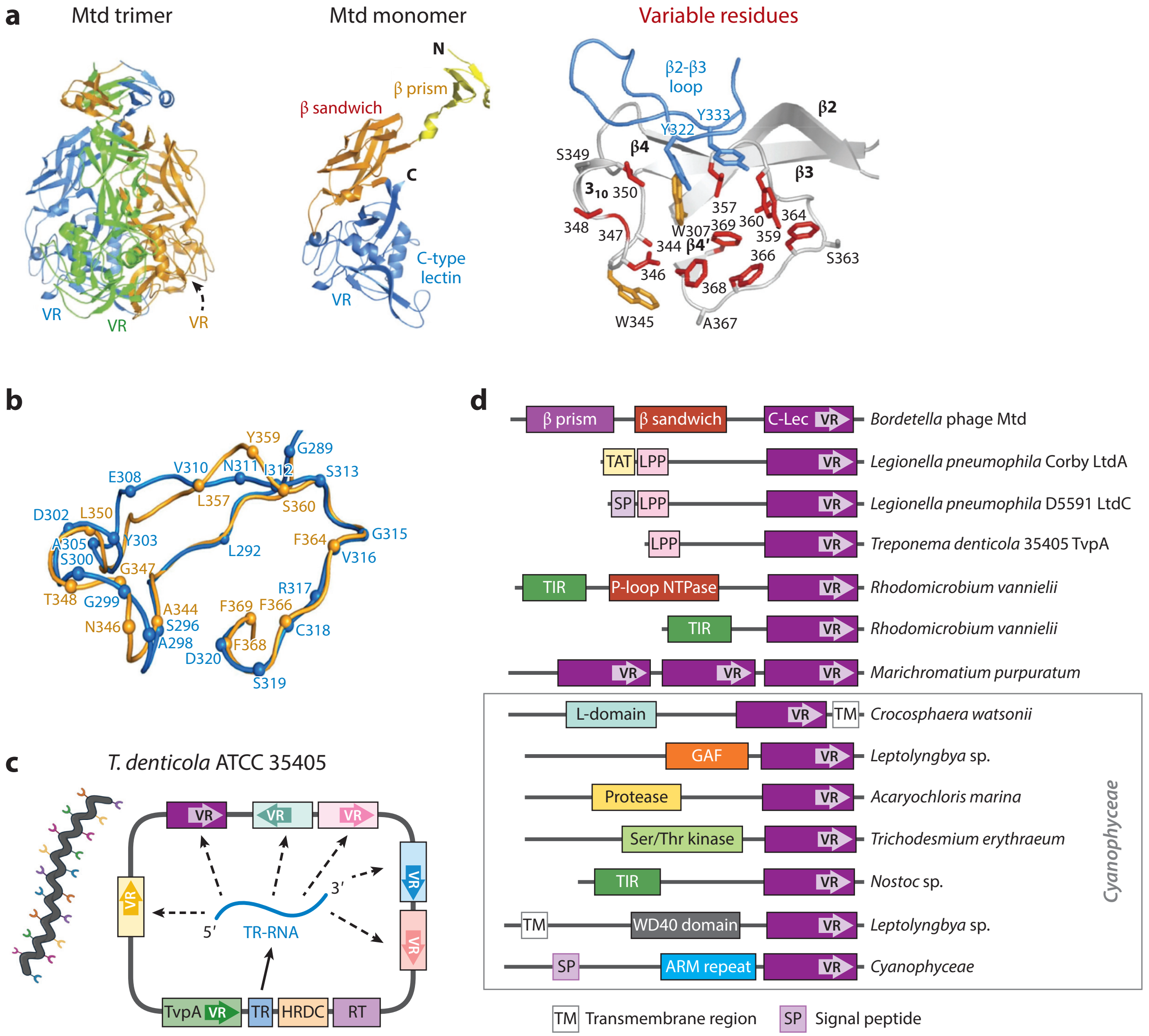
VARIABLE PROTEIN STRUCTURE AND FUNCTION

The selective advantage conferred by a DGR can only be appreciated in the context of the function of the diversified protein. Given the extensive heterogeneity of variable proteins, and the fact that nearly 70% lack functionally informative annotations other than the presence of a diversified VR (56, 86, 97), the physiological roles of the great majority remain elusive. Nonetheless, the assortment of motifs with predictable functions that have been identified in variable proteins indicates that DGR-directed mutagenesis has broad utility in the microbial world.

The most extensively characterized DGR variable protein, and one of the few whose precise function is known, is the BPP-1 tail fiber–associated receptor-binding protein, Mtd (23, 54, 55, 66, 68). Structural and functional studies of this prototypical DGR target reveal design principles that are likely conserved across variable proteins in general (**Figure 4*a***). Mtd directs the binding specificity of the phage, and diversification results from multiple nonsynonymous amino acid substitutions within its VR. These specifically occur at solvent-exposed residues within the binding pocket of a C-type lectin (C-Lec) fold (66). Selective diversification of the residues that make up the ligand-binding pocket drive the evolution of new Mtd specificities and can result in an exponential amplification of the binding strengths of these residues by avidity across all six tail fibers of the phage (68). Below, we discuss conserved structural features of Mtd and other characterized variable proteins in more detail before delving into the numerous types of predicted DGR targets that have been observed in microbial genomes.

Coevolution Between TR Sequence and Variable Protein Structure

Sequences that encode the TR-RNA template evolve to perform multiple functions. In BPP-1, the TR-RNA includes the polymerase-binding site and the initiating nucleotide for cDNA synthesis (37, 73). From a structural perspective, the positioning of adenines in a given TR coevolves with the location of codons in its cognate VR that encode amino acids that can participate in ligand-binding interactions and are able to tolerate mutagenesis without disrupting protein structure (**Figure 4*a*,*b***). From a sequence perspective, adenines are distributed in TRs in a manner that maximizes amino acid diversity. As initially shown for Mtd (**Figure 1*b***), TR adenines are most often positioned to diversify the first two nucleotides in VR codons, where random mutagenesis is most likely to lead to amino acid changes (97). Confirmation of this mutagenesis pattern came from longitudinally sampled human microbiomes where the diversity of individual DGR target genes was tracked over time. In these samples, mutagenic retrohoming was far more likely to produce nonsynonymous mutations than synonymous mutations, in contrast to other genes where synonymous mutations tended to dominate (86). One extraordinary TR sequence is the 5′-AAY-3′ motif, as adenine mutagenesis of cognate VR codons can generate a total of 15 amino acids that



(*Caption appears on following page*)

Figure 4 (*Figure appears on preceding page*)

Diverse functions of DGR-diversified variable proteins. (*a*) Crystal structure of the BPP-1 variable protein, Mtd. The pyramid-shaped Mtd trimer (*left*) contains three diversifiable VR regions located at the base of the complex, with two Mtd trimers located at the distal ends of each of the six BPP-1 tail fibers. The Mtd monomer (*middle*) comprises N-terminal (β-prism), intermediate (β-sandwich), and C-terminal (C-Lec) domains. On the right, solvent-exposed variable residues along the ligand-binding surface of Mtd are shown in red. (*b*) Superimposed VR regions of BPP-1 Mtd (*light orange*) and *Treponema denticola* TvpA (*blue*) demonstrating conservation of the C-Lec backbone structure across variable proteins that differ in primary sequence. Diversified residues are shown as spheres: 12 variable residues for Mtd and 20 for TvpA. Variable sites in Mtd show a remarkable degree of spatial correspondence with variable positions in TvpA, with additional diversified residues in TvpA interspersed between shared variable sites with minimal perturbation of the C-Lec secondary structure. (*c*) The *T. denticola* ATCC 35405 genome incudes a chromosomal DGR that encodes the TvpA variable protein, an accessory protein of unknown function (HDRC), and a DGR family reverse transcriptase. In addition to *tvpA*, six unlinked loci distributed throughout the *T. denticola* genome contain VR sequences at their 3′ ends that are nearly identical, except at sites corresponding to adenines in TR (52). This suggests that all seven genes are independently diversified by the same DGR-encoded machinery. The seven variable proteins are likely paralogs, with 25–67% amino acid identity, and several are predicted to have N-terminal lipoprotein localization signals that anchor them to the spirochete surface and position C-terminal variable residues for ligand binding. (*d*) Maps of DGR target genes and their functional domains. Example organisms are listed to the right of each variable protein. The modularity of DGR variable proteins is exemplified by similar C-Lec domains (*magenta*) with diversified VR sequences connected to a variety of N-terminal functional domains. With the exception of the first seven, all examples belong to the class *Cyanophyceae*. Abbreviations: ARM, armadillo repeat; BPP-1, Bvg-plus phage 1; C-Lec, C-type lectin; DGR, diversity-generating retroelement; GAF, cGMP-specific phosphodiesterase; LdtA, *Legionella* DGR target A; LPP, lipoprotein processing; Mtd, major tropism determinant; RT, reverse transcriptase; SP, signal peptide; TAT, twin arginine translocation; TIR, Toll/interleukin receptor; TM, transmembrane region; TR, template repeat; TvpA, *Treponema* variable protein A; VR, variable repeat. Panel *b* adapted from Reference 52. Panels *c* and *d* adapted from images created with BioRender.com.

range in charge, hydrophobicity, and size of their side chains but will never result in a stop codon (66, 69). By favoring the generation of nonsynonymous mutations, mutagenic retrohoming efficiently produces unique variable proteins and increases the likelihood of functionally diversifying DGR targets.

C-Type Lectin Folds Can Display Prodigious Amounts of Functional Diversity

TvpA is a DGR-diversified, surface-exposed lipoprotein encoded by the human oral spirochete *T. denticola* (25, 52). The discovery that TvpA displays variable residues using essentially the same C-Lec fold as Mtd, despite having only 16% local amino acid identity, provided the first structural evidence for the conservation of the C-Lec fold as a means to display massive sequence variation (52) (**Figure 4*b***). Although initially described as calcium-dependent carbohydrate-binding motifs, C-Lec folds are well-known to bind numerous different proteins, lipids, and small-molecule ligands, in addition to carbohydrates and polysaccharides (14, 22, 96). Classification of the DGR C-Lec fold can be difficult due to amino acid sequence heterogeneity and the lack of conserved disulfide bond–forming cysteines, as seen in other C-Lec motifs (66). Assignment, therefore, is most often based on homology comparisons through predicted three-dimensional structures (47).

For three DGR variable proteins with known structures, TvpA [Protein Data Bank (PDB) 2Y3C]; Mtd (PDB 1YU0); and a *Thermus aquaticus* variable protein, TaqVP (PDB 5VF4), diversified residues are displayed within the C-Lec binding pocket, and they are all solvent exposed, consistent with the hypothesis that DGRs generally function to evolve binding specificity (38, 52, 66, 68). Conversely, nondiversified amino acid side chains within the VR maintain the scaffold for the C-Lec fold and are more conserved than predicted (38, 66, 86, 97). This organization leads to a constrained and specific DGR diversification pattern where a subset of sites, tolerant of mutagenesis, are modified to alter ligand interactions while leaving residues that provide structural integrity to the C-Lec fold intact (**Figure 4*b***). In a recent survey of 32,321 DGR variable proteins predicted from diverse genomes and metagenomes, N-terminal domains varied widely, whereas VR-encoded sequences were primarily associated with C-Lec folds (86).

Mutagenic Retrohoming Evolves Avidity-Driven Interactions

Although the C-Lec fold motif directs binding specificity to particular ligands, avidity interactions conferred by multivalency can exponentially amplify individual binding strengths. This amplification relaxes the need for high affinity and optimal complementarity between interacting partners, thereby expanding the range of potential ligands that can be recognized by DGR-diversified proteins (66, 68). To illustrate the importance of avidity, we provide a vignette focused on tropism switching in a *Bordetella* phage (Vignette 1).

Vignette 1: DGR-directed protein evolution exploits avidity. To gain a better understanding of the receptor-ligand interactions driving infection by BPP-1, Miller et al. (68) cocrystalized the phage tail fiber protein Mtd (Mtd-P1) with its preferred ligand, the *Bordetella* outer membrane protein pertactin. Using surface plasmon resonance, these investigators found that the interaction between trimeric Mtd-P1 and pertactin was surprisingly weak ($K_d = 3.5\ \mu M$). After they selected for a DGR-mediated tropism switch, they isolated a descendant phage that expressed a diversified Mtd protein, Mtd-M1, that preferentially recognizes a Bvg$^-$ phase ligand for infection. The K_d value for trimeric Mtd-M1 and pertactin was estimated to be ~0.5 mM. While both Mtd variants had weak affinity toward pertactin, only phages with the Mtd-P1 variant could infect *Bordetella* using pertactin as a ligand. To explain these observations, knowledge of how the phage recognizes its target is required. Often measured between receptors and ligands, affinity is the strength of a biomolecular interaction; in the case of either Mtd-P1 or Mtd-M1 plus pertactin, it is quite low. Avidity, on the other hand, is measured by the accumulated strengths of all interactions within a complex, which include the affinity of the individual receptors, the total number of interactions (valency), and the structural rearrangements that occur to effectively position receptors (i.e., reduced dimensionality) (29). Each of the six BPP-1 tail fibers includes 2 Mtd trimers, giving 12 trimers or 36 Mtd monomers per phage (54, 55, 100). Binding interactions conferred by any of the tail fibers position the others close to the cell surface, thereby increasing the chances of an interaction with pertactin. Accordingly, the authors found that the strength of the binding interaction between Mtd-P1 and pertactin was amplified ~10^6-fold when measured in the context of the entire bacteriophage, while the weaker Mtd-M1 interaction was only amplified by ~10^3-fold, which explains the difference in selectivity. Multivalency and the differential amplification of binding strengths by avidity may relax the demand for optimal complementarity between Mtd and pertactin while enhancing distinctions among binding events that provide selectivity. This underlying principle of selective recognition is shared with immunoreceptors in the vertebrate immune system and is predicted to be a key feature of the evolution of variable protein repertoires by DGRs (68, 95).

DGR Variable Proteins Can Have Widely Diverse Functions

The majority of DGR target genes are predicted to encode modular, multidomain proteins composed of C-terminally positioned C-Lec folds that are diversified to evolve new ligand-binding interactions, appended to N-terminal domains that provide functional specificity to the protein (86, 95, 97). Although the majority lack informative annotations, several exceptions show that N-terminal regions of variable proteins can include a diverse range of functional domains, including peptidase, kinase, phosphodiesterase, hydrolase, and protein-protein interaction domains, among many others (**Figure 4*d***). This modular organization may promote rearrangements that swap N-terminal sequences to create novel chimeric proteins with rapidly evolvable ligand-binding capabilities. Additional information such as cellular localization can sometimes be predicted by the presence of transmembrane sequences, N-terminal lipobox motifs, or other secretion signals [e.g., Sec or twin arginine translocation (TAT)]. Indeed, the majority of bacterial DGR target genes that are not associated with phage genomes are predicted to encode variable proteins that localize to bacterial cell surfaces (86), as demonstrated in *L. pneumophila* (7, 52) and *T. denticola* (7, 52), although numerous exceptions exist (95). Finally, in addition to encoding at least one target gene subject to diversification, many DGRs appear to diversify remote target genes that are also subject to mutagenic retrohoming. These and other characteristics of variable protein repertoires are discussed in Vignette 2.

Vignette 2: Multiple remote DGR target genes encode variable cell surface lipoproteins in *Treponema denticola* and putative signaling proteins in cyanobacteria. A consequence of the mechanism of information flow during mutagenic retrohoming is that the TR and VR do not need to be in close proximity. As shown for BPP-1 phage (34) and a DGR carried on an ICE in *L. pneumophila* strain Corby (7), loci encoding a TR, an RT, and accessory proteins like Avd are able to efficiently diversify cognate VR sequences when expressed in *trans* on a replicating plasmid. Illustrating this mechanistic flexibility, the periodontal pathogen *T. denticola* appears capable of expressing multiple variable lipoproteins on its surface, a subset of which are encoded by unlinked genes distributed throughout the genome that are independently diversified by the same cognate TR (52) (**Figure 4*c***).

To imagine the benefits *T. denticola* might gain by encoding a variable surface protein array, it helps to consider the ecological context. The oral microbiota comprises hundreds of bacterial species and is thought to partly rely on coaggregation, or specific recognition of genetically distinct cells, to properly form (1, 50, 62, 63). Coaggregation is mediated through specific ligand-receptor binding interactions between distinct organisms that occur in a successive fashion (48, 85). Early colonizers of the oral cavity, which are usually dominated by gram-positive cocci such as *Streptococcus mitis* and *Streptococcus oralis* (53, 78) but can also include diverse members of the *Actinomyces*, *Gemella*, *Granulicatella*, *Neisseria*, *Prevotella*, *Rothia*, and *Veillonella* genera, bind to the tooth surface pellicle and to each other. Next, additional rounds of oral bacteria, eventually including *T. denticola*, bind to surface molecules displayed by earlier colonizers and to each other as they form complex microbial communities. Since the availability of potential ligands for *T. denticola* binding may vary greatly depending on community composition, which is highly dynamic (62), the ability to rapidly evolve ligand binding specificities of surface-displayed variable proteins could facilitate colonization (52). A potential advantage of expressing multiple variable protein genes that are independently diversified is that new biding specificities can evolve without the necessary loss of all previously existing interactions.

In contrast to surface-displayed variable proteins in *T. denticola*, *L. pneumophila*, and other bacteria, many DGR targets in diverse cyanobacterial species lack recognizable secretion signals and are predicted to be intracellular, while others have both extracellular and intracellular domains (95). A wide array of motifs, with predicted activities that range from protein-protein interactions to protease, hydrolase, NTPase, kinase, and other enzymatic functions, are appended to diversified C-terminal domains (**Figure 4*b***). Serine/threonine kinase (STK) motifs are particularly well represented among cyanobacterial variable proteins (95), and as in *T. denticola* they are often encoded at remote genomic sites distant from the cognate DGR.

STKs control complex processes in cyanobacteria, including developmental programs, adaptations to stress, and pathogenicity (95). Speculating on the possible selective advantages of variability, if diversified C-terminal domains exert control over kinase activity, through autoinhibition for example, mutagenic retrohoming could change the nature of signals that induce or repress activity. Alternatively, C-terminal domains could determine substrate specificity through direct binding interactions, or occupancy in distinct signaling complexes. Functional assessments challenging these and other hypotheses are eagerly awaited. DGR-mediated diversification of ligand-binding domains of STKs could confer a level of plasticity to signal transduction networks that has yet to be explored.

QUESTIONS AND APPLICATIONS

Much has been learned since DGRs were first discovered, yet many key features of these unique retroelements remain mysterious, and in some cases perplexingly so. From a mechanistic perspective, perhaps the greatest mystery is how, exactly, selective infidelity is achieved by DGR RTs. The advantage of this capability is obvious, as it allows diversified amino acids to be strategically placed within an otherwise stable scaffold for functional display, but the underlying biophysical mechanism is anything but obvious. The fact that adenines are universally associated with sites of mutagenesis suggests their recognition may somehow trigger a temporary relaxation of fidelity during reverse transcription (39), but the nature of the fidelity switch is unknown. A second gap is understanding how the resulting cDNA integrates into the VR and replaces the parental allele with an adenine-mutagenized copy. *cis*-Acting signals that flank VR sequences in target genes play an important role (35, 73), but little information is available beyond that. If integration involves strand displacement, the question of how assimilation of cDNA with such a high density of mismatches interfaces with host encoded DNA repair systems becomes relevant. Host factors may also be important for other events, including the integration step itself, but their identities, their functions, and the extent to which they determine DGR host range or control activity are undefined.

Accelerated evolution may benefit organisms that are far from their phenotypic optima, but unchecked mutagenesis can also increase the potential for loss of fitness, especially in stable environments. Since adenine mutagenesis is stochastic and loss-of-function mutations likely predominate over adaptive ones (25), it seems reasonable to predict that mutagenic retrohoming will be subject to regulation in many cases. In support of this hypothesis, DGR activity varies considerably across different environments (86). Under laboratory conditions, or in bioreactors where nutrients are abundant and cells are relatively unstressed, DGR activity is often low. Conversely, in competitive environments such as the intestinal microbiota, DGR-mediated diversification is prevalent. The benefits of uncovering regulatory mechanisms that modulate mutagenic retrohoming activity include a better understanding of not only the biology of DGRs but also the ability to effectively harness their activities.

Given the number of phylogenetically distant organisms with DGRs and the diversity of the elements themselves, there is much to be discovered regarding the functions of variable proteins and variable protein repertoires. Identifying preferred ligands and determining the consequences of ligand binding will be key. For variable proteins displayed on cell surfaces, ligand binding could facilitate nutrient acquisition, interactions with hosts or host products, colonization of environmental surfaces, or numerous other phenotypes. For cytosolic variable proteins, it will be interesting to understand how ligand binding to VR sequences affects the activities of N-terminal enzymatic or binding motifs. Bioinformatic comparisons will become increasingly useful for predicting potential functions (8, 46, 82), but they are unlikely to substitute for direct assessments in appropriate experimental systems.

Finally, DGR-directed mutagenesis is programmable and site specific, and it occurs in the absence of the kind of external manipulations required for phage-based or other protein display methods (5, 44). Because the TR is unmodified during mutagenic retrohoming, diversification can be repeated ad infinitum. DGRs can be engineered to diversify heterologous sequences, as was demonstrated using a reporter assay that restores kanamycin resistance through mutagenic retrohoming (73). As DGRs become better characterized, they will provide an increasingly useful tool kit for a range of bioengineering applications.

DISCLOSURE STATEMENT

J.F.M. is a cofounder of Pylum Biosciences, a company dedicated to developing antibacterial therapies that are precise and effective and have programmable targeting specificity.

ACKNOWLEDGMENTS

B.R.M. gratefully acknowledges research fellowship support from the UCLA–NIH Microbial Pathogenesis Training Program (5 T32 AI007323–30) and career development support from the Pediatric Scientist Development Program (5 K12 HD000850–38). J.F.M. acknowledges and thanks the Kavli Foundation, the Shaffer Family Foundation, and the California NanoSystems Institute for their support.

LITERATURE CITED

1. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 43:5721–32
2. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J. Virol.* 84:9864–78
3. Ackermann HW. 1998. Tailed bacteriophages: the order Caudovirales. *Adv. Virus Res.* 51:135–201
4. Alayyoubi M, Guo H, Dey S, Golnazarian T, Brooks GA, et al. 2013. Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase. *Structure* 21:266–76
5. Alfaleh MA, Alsaab HO, Mahmoud AB, Alkayyal AA, Jones ML, et al. 2020. Phage display derived monoclonal antibodies: from bench to bedside. *Front. Immunol.* 11:1986
6. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7:13219
7. Arambula D, Wong W, Medhekar BA, Guo H, Gingery M, et al. 2013. Surface display of a massively variable lipoprotein by a *Legionella* diversity-generating retroelement. *PNAS* 110:8212–17
8. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373:871–76
9. Belcher T, Dubois V, Rivera-Millot A, Locht C, Jacob-Dubuisson F. 2021. Pathogenicity and virulence of *Bordetella pertussis* and its adaptation to its strictly human host. *Virulence* 12:2608–32
10. Benler S, Cobian-Guemes AG, McNair K, Hung SH, Levi K, et al. 2018. A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* 6:191
11. Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* 11:14–28
12. Bondy-Denomy J, Davidson AR. 2014. When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *J. Microbiol.* 52:235–42
13. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, et al. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523:208–11
14. Brown GD, Willment JA, Whitehead L. 2018. C-type lectins in immunity and homeostasis. *Nat. Rev. Immunol.* 18:374–89
15. Brussow H, Canchaya C, Hardt WD. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68:560–602
16. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. *Cell* 184:1098–109.e9
17. Campillo-Balderas JA, Lazcano A, Becerra A. 2015. Viral genome size distribution does not correlate with the antiquity of the host lineages. *Front. Ecol. Evol.* 3. **https://doi.org/10.3389/fevo.2015.00143**
18. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, et al. 2015. Genomic expansion of domain Archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* 25:690–701

19. Chi X, Li Y, Qiu X. 2020. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology* 160:233–47
20. Christensen SM, Eickbush TH. 2005. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol. Cell. Biol.* 25:6617–28
21. Correa AMS, Howard-Varona C, Coy SR, Buchan A, Sullivan MB, Weitz JS. 2021. Revisiting the rules of life for viruses of microorganisms. *Nat. Rev. Microbiol.* 19:501–13
22. Cummings RD, McEver RP. 2009. C-type lectins. In *Essentials of Glycobiology*, ed. A Varki, RD Cummings, JD Esko, HH Freeze, P Stanley, et al., pp. 459–74. New York: Cold Spring Harb.
23. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. 2010. Three-dimensional structure of tropism-switching *Bordetella* bacteriophage. *PNAS* 107:4347–52
24. Derr LK, Strathern JN. 1993. A role for reverse transcripts in gene conversion. *Nature* 361:170–73
25. 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* 431:476–81
26. Duffy S. 2018. Why are RNA virus mutation rates so damn high? *PLOS Biol.* 16:e3000003
27. Duffy S, Holmes EC. 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J. Virol.* 82:957–65
28. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* 27:140–53.e9
29. Erlendsson S, Teilum K. 2020. Binding revisited—avidity in cellular function and signaling. *Front. Mol. Biosci.* 7:615565
30. Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* 24:363–67
31. Gong J, Qing Y, Guo X, Warren A. 2014. "*Candidatus* Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). *Syst. Appl. Microbiol.* 37:35–41
32. Griffiths AJF. 2012. *Introduction to Genetic Analysis*. New York: W.H. Freeman
33. Guo H, Arambula D, Ghosh P, Miller JF. 2014. Diversity-generating retroelements in phage and bacterial genomes. *Microbiol. Spectr.* 2. **https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014**
34. Guo H, Tse LV, Barbalat R, Sivaamnuaiphorn S, Xu M, et al. 2008. Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Mol. Cell* 31:813–23
35. Guo H, Tse LV, Nieh AW, Czornyj E, Williams S, et al. 2011. Target site recognition by a diversity-generating retroelement. *PLOS Genet.* 7:e1002414
36. Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. *Annu. Rev. Ecol. Evol. Systemat.* 40:151–72
37. Handa S, Jiang Y, Tao S, Foreman R, Schinazi RF, et al. 2018. Template-assisted synthesis of adenine-mutagenized cDNA by a retroelement protein complex. *Nucleic Acids Res.* 46:9711–25
38. Handa S, Paul BG, Miller JF, Valentine DL, Ghosh P. 2016. Conservation of the C-type lectin fold for accommodating massive sequence variation in archaeal diversity-generating retroelements. *BMC Struct. Biol.* 16:13
39. Handa S, Reyna A, Wiryaman T, Ghosh P. 2021. Determinants of adenine-mutagenesis in diversity-generating retroelements. *Nucleic Acids Res.* 49:1033–45
40. Hausner G, Hafez M, Edgell DR. 2014. Bacterial group I introns: mobile RNA catalysts. *Mob. DNA* 5:8
41. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, et al. 2016. A new view of the tree of life. *Nat. Microbiol.* 1:16048
42. Ives AR, Garland T Jr. 2010. Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* 59:9–26
43. Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. *Annu. Rev. Microbiol.* 56:489–520
44. Jaroszewicz W, Morcinek-Orłowska J, Pierzynowska K, Gaffke L, Węgrzyn G. 2022. Phage display and other peptide display technologies. *FEMS Microbiol. Rev.* 46(2):fuab052

45. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* 299:27–51

46. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–89

47. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10:845–58

48. Kolenbrander PE, Palmer RJ Jr., Rickard AH, Jakubovics NS, Chalmers NI, Diaz PI. 2006. Bacterial interactions and successions during plaque development. *Periodontol. 2000* 42:47–79

49. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* 31:147–57

50. Kuramitsu HK, He X, Lux R, Anderson MH, Shi W. 2007. Interspecies interactions within oral microbial communities. *Microbiol. Mol. Biol. Rev.* 71:653–70

51. Lautner M, Schunder E, Herrmann V, Heuner K. 2013. Regulation, integrase-dependent excision, and horizontal transfer of genomic islands in *Legionella pneumophila*. *J. Bacteriol.* 195:1583–97

52. Le Coq J, Ghosh P. 2011. Conservation of the C-type lectin fold for massive sequence variation in a *Treponema* diversity-generating retroelement. *PNAS* 108:14649–53

53. Li J, Helmerhorst EJ, Leone CW, Troxler RF, Yaskell T, et al. 2004. Identification of early microbial colonizers in human dental biofilm. *J. Appl. Microbiol.* 97:1311–18

54. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, et al. 2002. Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* 295:2091–94

55. 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.* 186:1503–17

56. Lobb B, Tremblay BJ, Moreno-Hagelsieb G, Doxey AC. 2020. An assessment of genome annotation coverage across the bacterial tree of life. *Microb. Genom.* 6:e000341

57. Loewe L, Hill WG. 2010. The population genetics of mutations: good, bad and indifferent. *Philos. Trans. R. Soc. Lond. B* 365:1153–67

58. 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

59. Ma L, Dewan KK, Taylor-Mulneix DL, Wagner SM, Linz B, et al. 2021. Pertactin contributes to shedding and transmission of *Bordetella bronchiseptica*. *PLOS Pathog.* 17:e1009735

60. Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* 16:793–805

61. Maniloff J, Ackermann HW. 1998. Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. *Arch. Virol.* 143:2051–63

62. Mark Welch JL, Ramirez-Puebla ST, Borisy GG. 2020. Oral microbiome geography: micron-scale habitat and niche. *Cell Host Microbe* 28:160–68

63. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. 2016. Biogeography of a human oral microbiome at the micron scale. *PNAS* 113:E791–800

64. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, et al. 1998. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* 188:2151–62

65. Mazel D. 2006. Integrons: agents of bacterial evolution. *Nat. Rev. Microbiol.* 4:608–20

66. 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.* 12:886–92

67. Melvin JA, Scheller EV, Miller JF, Cotter PA. 2014. *Bordetella pertussis* pathogenesis: current and future challenges. *Nat. Rev. Microbiol.* 12:274–88

68. Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, Ghosh P. 2008. Selective ligand recognition by a diversity-generating retroelement variable protein. *PLOS Biol.* 6:e131

69. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD. 2012. Hypervariable loci in the human gut virome. *PNAS* 109:3962–66

70. Moxon R, Bayliss C, Hood D. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* 40:307–33

71. Muller F, Tobler H. 2000. Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parascaris univalens*. *Int. J. Parasitol.* 30:391–99

72. Murphy K, Weaver C, Janeway C. 2017. *Janeway's Immunobiology*. New York: Garland Sci.

73. Naorem SS, Han J, Wang S, Lee WR, Heng X, et al. 2017. DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. *PNAS* 114:E10187–95

74. Nayfach S, Paez-Espino D, Call L, Low SJ, Sberro H, et al. 2021. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* 6:960–70

75. Nelson WC, Stegen JC. 2015. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front. Microbiol.* 6:713

76. 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

77. Nishijima S, Suda W, Oshima K, Kim SW, Hirose Y, et al. 2016. The gut microbiome of healthy Japanese and its microbial and functional uniqueness. *DNA Res.* 23:125–33

78. Palmer RJ Jr., Gordon SM, Cisar JO, Kolenbrander PE. 2003. Coaggregation-mediated interactions of streptococci and actinomyces detected in initial human dental plaque. *J. Bacteriol.* 185:3400–9

79. Papavasiliou FN, Schatz DG. 2002. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell* 109(Suppl):S35–44

80. Paul BG, Bagby SC, Czornyj E, Arambula D, Handa S, et al. 2015. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nat. Commun.* 6:6585

81. Paul BG, Burstein D, Castelle CJ, Handa S, Arambula D, et al. 2017. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. *Nat. Microbiol.* 2:17045

82. Paul BG, Eren AM. 2022. Eco-evolutionary significance of domesticated retroelements in microbial genomes. *Mobile DNA* 13:6

83. Potapov V, Ong JL. 2017. Examining sources of error in PCR by single-molecule sequencing. *PLOS ONE* 12:e0169774

84. Rapid amplification of 5′ complementary DNA ends (5′ RACE). 2005. *Nat. Methods* 2:629–30

85. Rickard AH, Gilbert P, High NJ, Kolenbrander PE, Handley PS. 2003. Bacterial coaggregation: an integral process in the development of multi-species biofilms. *Trends Microbiol.* 11:94–100

86. Roux S, Paul BG, Bagby SC, Nayfach S, Allen MA, et al. 2021. Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* 12:3076

87. 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 Genom.* 13:430

88. Schultz SJ, Champoux JJ. 2008. RNase H activity: structure, specificity, and function in reverse transcription. *Virus Res.* 134:86–103

89. Sharifi F, Ye Y. 2019. MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res.* 47:W289–94

90. Stokes HW, Gillings MR. 2011. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* 35:790–819

91. Svensson EI, Berger D. 2019. The role of mutation bias in adaptive evolution. *Trends Ecol. Evol.* 34:422–34

92. Taylor VL, Fitzpatrick AD, Islam Z, Maxwell KL. 2019. The diverse impacts of phage morons on bacterial fitness and virulence. *Adv. Virus Res.* 103:1–31

93. Toro N, Nisa-Martinez R. 2014. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLOS ONE* 9:e114083

94. Trzilova D, Tamayo R. 2021. Site-specific recombination—how simple DNA inversions produce complex phenotypic heterogeneity in bacterial populations. *Trends Genet.* 37:59–72

95. Vallota-Eastman A, Arrington EC, Meeken S, Roux S, Dasari K, et al. 2020. Role of diversity-generating retroelements for regulatory pathway tuning in cyanobacteria. *BMC Genom.* 21:664

96. Weis WI, Taylor ME, Drickamer K. 1998. The C-type lectin superfamily in the immune system. *Immunol. Rev.* 163:19–34

97. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, et al. 2018. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res*. 46:11–24

98. Yan F, Yu X, Duan Z, Lu J, Jia B, et al. 2019. Discovery and characterization of the evolution, variation and functions of diversity-generating retroelements using thousands of genomes and metagenomes. *BMC Genom*. 20:595

99. Ye Y. 2014. Identification of diversity-generating retroelements in human microbiomes. *Int. J. Mol. Sci.* 15:14234–46

100. Zhang X, Guo H, Jin L, Czornyj E, Hodes A, et al. 2013. A new topology of the HK97-like fold revealed in *Bordetella* bacteriophage by cryoEM at 3.5 Å resolution. *eLife* 2:e01299

101. 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* 83:529–38

102. Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* 82:545–54