Title
Legionella Diversity Generating Retroelements: Creating massively variable repertoires of surface displayed proteins

Author
Arambula, Diego

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Legionella Diversity Generating Retroelements: Creating massively variable repertoires of surface displayed proteins

A dissertation partially satisfying the requirements for the degree

Doctor of Philosophy

in Microbiology, Immunology and Molecular Genetics

by

Diego Arambula

2014

ABSTRACT OF THE DISSERTATION

Legionella Diversity Generating Retroelements: Creating massively variable repertoires of surface displayed proteins

by

Diego Arambula

Doctor of Philosophy in Microbiology, Immunology and Molecular Genetics

University of California, Los Angeles, 2014

Professor Jeffery F. Miller, Chair

Diversity-generating retroelements (DGRs) are distinguished by their ability to iteratively diversify defined DNA sequences which encode the ligand binding domains of target proteins (TP). Diversification occurs through a template-dependent, error prone reverse-transcriptase mediated process, termed mutagenic homing, which introduces nucleotide substitutions into a variable repeat (VR) while preserving cis- and trans-acting elements needed for future rounds of diversification. The process of DNA diversification requires a DGR-encoded reverse transcriptase (RT), an accessory

variability determinant (Avd), and a template repeat (TR)-derived RNA intermediate. The archetype DGR is found within the *Bordetella* bacteriophage BPP however, over 300 putative DGRs have been identified within the bacterial domain as well as within a species of archaea. We have identified DGRs within the opportunistic human pathogen *Legionella pneumophila* (*Lp*) as well as within *Legionella tunisiensis* which encode nearly identical diversification machinery, yet each has a VR with a unique pattern of adenine mutagenesis suggesting individualized diversification in response to selection. We analyzed the DGR within *Lp* strain Corby and identified that the genetic requirements for mutagenic homing were similar to those of BPP, suggesting all DGRs might function through a conserved mechanism. Using *in vitro* growth conditions, we observed elevated levels of diversification during *Lp* transition from exponential to stationary growth phase suggesting that DGR mutagenic homing might be modulated by host regulatory networks. To investigate this hypothesis we generated deletions of key factors, *relA* and *spoT*, which are critical for coordinating phenotypic differentiation of *Lp*. PCR analysis to detect levels of mutagenic homing in *wt* and mutant *Lp* cells revealed an increase in *relA spoT* double deletion mutants. qRT-PCR analysis of *wt* and mutant cells showed that double deletion mutants had an increase in TR-RNA transcripts, while expression of *avd* and *RT* was unchanged. Over-expression of TR-RNA increased levels of mutagenic homing above those observed in *wt* cells, but not to the extent observed with the overexpression of *avd*, TR, and *RT*. Cumulatively these data suggest that in *Lp*, the abundance of TR-RNA transcripts partially controls rates of mutagenic homing, indicating regulation of these elements on multiple levels. *Lp* DGRs have the potential to generate ~10^{19} distinct polypeptide sequences within C-type lectin

(CLec) domains of their TPs. The *Lp* Corby TP, *ldtA*, expresses a surface displayed outer membrane (OM) lipoprotein whose CLec domain is exposed to the extra-cellular milieu. Translocation of LdtA across the inner membrane requires the twin-arginine translocation (TAT) machinery where we hypothesize it is recognized, modified, and transported to the outer membrane by the localization of lipoproteins (Lol) system. To fully understand the pathways required for surface display of a protein, we identified and analyzed the contribution of a non-canonical lipobox with conserved targeting residues at +2/+3 positions in LdtA. Mutagenesis of the lipobox conserved cysteine as well as replacement of targeting residues with amino acids shown to result in sorting by Lol to the OM all resulted in retention of LdtA in the inner membrane. Furthermore, cleavage of LdtA from pro- to mature-peptide was found to depend on the characteristic of the +2 residue suggesting that maturation and subsequent translocation of LdtA does not follow the established convention for lipoproteins. Furthermore, we demonstrated surface display of LdtA in several species of gram negative bacteria suggesting its localization requires conserved pathways common to bacteria. These results suggest that DGR TPs in *Lp* are diversified in response to the physiological state of the host and trafficked to the surface by an unusual Lol-related mechanism.

The dissertation of Diego Arambula has been approved.

Peter J. Bradley

Wenyuan Shi

James A. Wohlschlegel

Jeffery F. Miller, Committee Chair

University of California, Los Angeles

2014

...for those who came before me, for those who helped me, and for those who will come after me…

Table of Contents

Acknowledgments

I have more people to thank than space to thank them in. I will always be indebted to my mentor Dr. Jeff F. Miller, who took a chance on me and probably spent more time on me than he should have but has shaped me into the scientist I am today. To the members of the laboratory, both past and present, who are like an extended family specifically, I would like to thank Bob and Ruchi Medhekar, Asher Hodges, Dave Richards, and Ming Liu who were there in the beginning. As well as to Atish Ganguly, Umesh Ahuja, Isabelle Spears, Elizabeth Czornyj, and especially Todd French for being there at the end. A special thanks to Huatao Guo, a scientific compatriot but more importantly he, along with a few others, taught me what it takes to ask the right question. This work would not have been possible without Mari Gingery (UCLA) and Steve Zimmerly (U. Calgary) from whom I shamelessly borrow the algorithm to identify DGRs and the comparison of DGRs to known families of retroelements, respectively.

On a personal note, I would like to thank my family and specifically my wife for they have supported me through the good as well as the terrible, terrible times. And last but certainly not least and mostly just because, Derrick West, Victor Ratliff, Jeramie Campbell, and Ernie Najera.

Biographical Sketch

I was born to a loving but mischievous home and spent time living in between Mexico and Arizona. I spent my youth learning to separate what must be done from what should be done. I received my bachelors of science in Biology from Arizona State University which was followed by an internship at the Veterans Affairs Boston Healthcare System. After working for Harvard University, I travelled across the country to attend graduate school at University of California Los Angeles.

Chapter 1. A brief introduction to retroelements and the opportunistic human pathogen *Legionella pneumophila*.

“Nature is always adapting to changing conditions and seeking equilibrium. Everything has a purpose, nothing is lost, nothing is wasted, and nothing is extraneous”

Mutations generate diversity which is selected for

Mutation is considered the force that generates nucleotide diversity in genes and this genetic variability leads to phenotypes or individuals within a population which natural selection acts upon to determine the fittest. While most mutations are considered to be neutral or deleterious, there are instances where increased mutation rates are beneficial to host fitness by facilitating adaptation to stressful environments [1]. These beneficial mutations of existing genes are generally considered to lead to refinement of function or generate novel function and the dramatic effect of single nucleotide polymorphisms (SNP) on host physiology has been described [2]. Microorganisms often face rapidly changing environments and their ability to quickly adapt has been contributed to an increased mutation rate, as well as the acquisition of new genes into or deletion of genes from their genomes [3]. The acquisition of genes, leading to an increased fitness, is especially poignant during host-pathogen interactions, exposure to antibiotics, and other non-lethal selections [1, 2, 4, 5]. Transposable elements (TEs) are a broad class of elements widely disseminated throughout microbial populations that are thought to be instrumental to the ability of microorganisms to adapt to stressful environments [6]. TEs alter host genomes through a variety of means and mechanisms including, but not limited to, modulating gene regulation, gene duplications, gene insertion, and chromosomal rearrangements [6].

Retroelements, a class of Transposable Elements

TEs, or mobile genetic elements, were first identified within the genome of *Zea mays* and have since been found within the chromosomes of almost every living organism. These elements duplicate themselves to move into new genomic locations and have been divided into two major classes based on their mechanism of mobility. Class II transposable elements typically use a gene product, except for miniature inverted repeats (MITEs), to facilitate their duplication and dissemination in a host genome. Unlike Class I transposable elements, they do not require an RNA intermediate, instead they use a cut-and-paste or rolling circle mechanism to excise from and integrate into DNA sequences. This class includes DNA transposons, MITEs, Helitrons, and Mavericks [6, 7]. Class I transposable elements are commonly called retroelements since they use a reverse transcriptase generated RNA-intermediate to copy-and-paste themselves into a new genomic location [7]. Retroelements within this class are further distinguished from each other based on the presence of flanking long terminal repeats (LTRs), such as retrotransposons and endogenous retroviruses, or the absence of LTRs (non-LTRs), such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and processed pseudogenes [8]. Group II introns are a class of mobile genetic elements which are thought to be an evolutionary ancestor of non-LTR retrotransposons that are found in bacteria, archaea, and some eukaryotic organelles. These elements encode a self-splicing RNA as well as a multifunctional reverse transcriptase which are necessary for its propagation and dissemination throughout the host genome [9, 10]. Retrons are a class of elements only

found in bacteria whose reverse transcriptase is similar to that of Group II Introns. These elements encode a small satellite DNA (msDNA) which replicates itself and jumps into new locations throughout the host genome [11]. Diversity-generating retroelements (DGRs) are a recently discovered family of retroelements which are related to bacterial Group II introns and retrons (Figure 1) [12]. Interestingly, while DGRs appear to have been derived from Group II introns, with an apparently similar mechanism of action, their contribution to host fitness appear to be vastly different.

Transposable elements-mechanisms of mobility

We will briefly discuss the mechanism of action for a select number of retroelements as points of comparisons for the DGR mutagenic homing.

Non-LTR Retroelements. LINEs are a well-studied class of non-LTR retrotransposons which direct their mobility through an RNA intermediate. The human LINE, L1, is composed of a 5'-untranslated region (UTR) followed by two open reading frames (ORFs) then a short UTR and a poly-A tail region. The 5' UTR functions as a promoter for the two ORFs, with the first ORF (ORF1) encoding a protein shown to bind nucleic acids and the second ORF (ORF2) encoding a dual purpose protein that serves as a endonuclease and reverse transcriptase [7]. Expression of L1 results in L1-RNA bound to the two protein products from ORF1/2 forming a ribonucleoprotein (RNP) complex. The L1 RNP then inserts into a new genomic location, a process call retrotransposition, through a mechanism called target primed reverse transcription (TPRT) where a DNA consensus motif is nicked to expose a single stranded nucleotide stretch that is used to form a DNA/RNA complex between the host chromosome and the

L1-RNA. This complex functions as a primer for reverse transcription using the L1 ORF2 protein. A second chromosomal nick exposes a hydroxyl group which is used for second strand synthesis. These two nicks are sealed and newly synthesized L1 DNA is integrated into the chromosome [13].

Group II Introns. Group II introns are mobile genetic elements that are comprised of a catalytic RNA coupled with an intron encoded protein (IEP) which has reverse transcriptase activity and is related to reverse transcriptases of non-LTR retrotransposons [10]. The catalytic RNA is auto-spliced via two trans-esterification steps into a matured RNA lariat which, along with the IEP, forms the RNP complex [9]. The RNP complex allows for Group II introns to proliferate throughout the host chromosome by facilitating invasion into eptopic sites [10]. Group II intron invasion involves RNP complex binding to the target site where the RNA lariat, using an exposed hydroxyl, reverse-splices into the DNA strand and serves as a template for IEP reverse transcription/synthesis of cDNA[9, 10].

Retrons. Retrons may be a mobile genetic element that is composed of at least three ORFs driven by a single promoter. The first two ORFs encode *msr* and *msd* while the third encodes *ret*, a reverse transcriptase related to viral elements [11]. These three genes are necessary for the production of msDNA which is composed of a single strand of DNA bound to a single strand of RNA and this DNA/RNA complex can be bound to retron proteins. While the function of msDNA is currently unknown, they are often produced in large numbers, up to hundreds of copies per cell. Retrons are found sporadically distributed throughout the bacterial domain and, in those phyla where they

are found, it is typically only found in a few species. The mobility of retrons is uncertain. While they are usually only found as a single copy per chromosome, instances have been reported of genomes with multiple partial copies of retrons that appear to have been duplicated by reverse transcription, suggesting these elements may be inducing their mobility [11].

Diversity-generating retroelements. DGRs are a recently discovered clade of retroelements whose reverse transcriptase (RT) is related to those found in Group II introns and retrons [12, 14]. All DGRs encode a target protein (*TP*), an accessory gene (*avd* or analogous gene), and a dedicated *RT*. These factors act in concert to iteratively diversify defined DNA sequences which encode the ligand biding domains of TPs [7, 15]. DNA diversification occurs through a template-dependent, error prone reverse-transcriptase mediated process, termed mutagenic homing, which introduces nucleotide substitutions while preserving cis- and trans- acting elements needed for future rounds of diversification. Furthermore, this transfer of sequence information is unidirectional and confined to a well-defined nucleotide region called the variable repeat (VR). Nucleotide variability is generated through a copy-diversify-and-replace mechanism that has been proposed to be a form of TPRT, similar to related retroelements (Figure 2) [16]. Interestingly, while the activity of most retroelements results in their duplication and distribution throughout the host genome, DGR activity results in diversification of VR, the introduction of polypeptide variability into TP, and ultimately results in an accelerated evolution of TP ligand binding domains. Mechanistic studies of the archetype DGR, the *Bordetella bronchiseptica* (*Bb*) bacteriophage (BPP), have revealed key observations about the precise mechanism by which mutagenic homing generates

diversity in target genes which will be discussed in detail below [15-19]. The study of the molecular mechanisms of this family of elements will be beneficial in understanding how closely related retroelements, with apparently similar modes of action, affect host fitness in disparate ways.

Diversity Generating Retroelements

The mechanism of DGR mutagenic homing.

Studies of mutagenic homing in BPP identified the requirement of the DGR-encoded RT, an accessory tropism determinant (*atd*), and a TR-derived RNA intermediate. This TR-RNA provides a template for reverse transcription and, during cDNA synthesis, TR adenine residues are copied into any of the four deoxyribonucleotides [16, 19]. The diversified cDNA displaces a VR found at the 3’ end of the TP encoding target gene [16]. The transfer of sequence information is a unidirectional, mutagenic retrotransposition process that is targeted to the VR. DGR mutagenic activity of DNA sequences is constrained through two cis-acting features found at the 3' end of VR; the initiation of mutagenic homing (IMH) element and a DNA hairpin/cruciform structure[16, 18, 19]. Structural studies of BPP show that TR adenines are precisely positioned to correspond to residues in the ligand binding pocket of the VR containing C-type lectin (CLec) domain at the C-terminus of the tail fiber protein, reflecting co-evolution between the genetic mechanism that generates diversity and the protein scaffold that displays it [18, 19].

DGR mutagenic homing occurs through an RNA intermediate. The essential role of the BPP RT suggested that DGR mutagenic homing occurs through RNA and not a DNA intermediate [17]. This was confirmed through intron-tagging of the TR with a self-splicing Group I intron followed by observation of transfer of sequence information. The resulting diversified progeny VR were found to contain ligated exons of the group I intron [16]. As the intron splices only during processing of an RNA molecule, this observation suggested that homing had occurred through and required synthesis of a TR-RNA intermediate.

The TR, and flanking nucleotide sequences, was investigated to determine boundaries for synthesis of the TR-RNA intermediate during mutagenic homing. Deletion analysis showed that, while most of the TR internal sequence is not essential, DGR homing absolutely requires the 5', 3', as well as flanking upstream and downstream sequences. Recently, it was shown that most of the 3' end of the *atd* ORF, which is found immediately upstream of TR, is an important part of the TR-containing RNA with deletion inhibiting mutagenic homing [20]. This suggests coevolution of the RNA intermediate and this essential protein. Additionally, we have identified that the synthesized TR-RNA intermediate includes downstream sequences which may overlap with the start codon of RT. Based on these observations, it has been predicted that the TR-RNA is presented as a specific RNA secondary or tertiary structure which may be conserved among DGRs. This TR-RNA intermediate is a template for RT and the resulting cDNA contains random nucleotides at positions corresponding to adenines in the TR (see below).

3’ Target recognition is both sequence- and structure-dependent for many DGRs. While DGRs have the potential to generate vast amounts of nucleotide diversity, in order for this diversity to be beneficial there must be tight constraints on what a target sequence is, as mutation of random nucleotide sequences would likely be deleterious to the host. For the BPP DGR, several key studies have been made on how the diversified cDNA integrates into or replaces, a process called target site recognition, the parental non-diversified chromosomal VR sequence. Deletion analysis demonstrated that target site recognition required sequences found in the VR as well as a downstream 24 base pair (bp) sequence. This 24 bp sequence found at the 3’ end of VR, which differs from a similar sequence at the 3’ of TR at five discrete sites, was found to determine the directionality of sequence information transfer during mutagenic homing. When the VR sequence was swapped for that of TR, the VR was no longer diversified. Replacing the sequence 3’ of TR with that of VR enabled the modified TR to be diversified [16]. Thus, this element in VR was named Initiation of Mutagenic Homing (IMH), while the corresponding element at the end of TR was called IMH*.

The 24 bp sequence downstream of VR includes two 8 bp GC-rich inverted repeats separated by a four nucleotide (nt) spacer which were predicted to form a DNA hairpin/cruciform structure with a four nt loop. Analysis where the 8 bp nucleotide sequence was changed to disrupt the potential DNA hairpin structure and then complemented showed that it was the DNA structure, instead of the primary sequence, that is important for target site recognition during mutagenic homing [19]. Further analysis of the BPP DNA stem loop structure showed that both the length and the GC content of the stem were important. The loop appeared to be especially critical as any

changes in sequence or size dramatically affected levels of homing. The absolute position of the stem-loop structure relative to VR is also important as short insertions were well tolerated however, longer insertions or deletions significantly inhibited mutagenic homing [19].

In silico analysis of putative DGRs found within nucleotide databases identified similar DNA hairpin/cruciform structures, containing 7-10 bp GC-rich stems and 4 nt in many phage genomes [21]. In addition to phage DGRs, stem loop structures were identified in a significant number of bacterial chromosomal DGRs suggesting DNA hairpin/cruciform structures have a conserved function, likely similar to that found in BPP however their exact role in mutagenic homing remains to be determined. Although potential stem loop structures are found in many DGRs, they are by no means ubiquitous and it is unclear if DNA sequences downstream of the corresponding VRs either adopt different structures or if the 3'-end target gene recognition in these DGRs occurs through alternative mechanisms.

A target-primed reverse transcription model for 3' cDNA integration. DGR mutagenic homing has been proposed to occur through a target DNA-primed reverse transcription, TPRT, mechanism. A TPRT model was initially demonstrated in the *Bombyx mori* R2 (R2Bm) element which is a site-specific retroelement lacking long terminal repeats. The RT encoded by the R2 element also functions as a DNA endonuclease nicking the DNA antisense strand shortly downstream of the R2 insertion site and uses the exposed sequence as a primer to reverse transcribe the R2 RNA. As a consequence, the cDNA is directly attached to the target DNA at the 3' end which is

essential for R2 transposition. A similar TPRT mechanism has been described in the mobility of group II introns and LINEs, two retroelements related to R2Bm and DGRs.

As DGR RTs are thought to have derived from group II introns and are related to LINE elements, it was likely that mutagenic homing functions through TPRT. Consistent with this, BPP DGR mutagenic homing was shown, similar to the retrohoming of bacterial group II introns, to be independent of host RecA-based recombination machinery. Additionally, it was observed that nucleotide polymorphisms at the 3' end of BPP TR (IMH*) were never transferred to the corresponding region of VR. Analysis of sequence information transferred during mutagenic homing using a marker transfer assay where markers were introduced in the tagged donor TR revealed a sharp marker co-conversion boundary within the IMH element of VR [19]. This observation lends additional support for the TPRT model, where 3' cDNA integration could occur in the absence of 5'-end cDNA integration. However, while TPRT is currently the prevailing model to explain DGR mutagenic homing alternative mechanisms may exist.

cDNA integration at the 5' end is short homology-mediated. As mentioned above, the only requirement for 3' cDNA integration is DNA sequences with homology to the VR IMH. However, 5' cDNA target site integration appears to require short stretches of homology that is sequence independent and does not require cDNA extension to the 5' terminus of TR [16]. This was demonstrated by inserting a short, homologous *mtd* sequence upstream of VR into an internally-deleted TR. DGR mutagenic homing with 5' cDNA integration was observed and sequence analysis revealed that 5' cDNA integration had occurred through short homologies between the

engineered TR and the target VR sequence, possibly via cDNA template switching or strand displacement. In the marker coconversion assay, single nucleotide markers introduced upstream of the tag in TR were transferred at varying frequencies suggesting that the elongating cDNA products could integrate before extending to the very 5' of TR. This is consistent with the observation of polar patterns of marker transfer in phage tropism switching assays.

Mechanism of DGR adenine-specific mutagenesis. Adenine-specific mutagenesis is a distinctive feature of DGRs but its mechanism remains largely unknown. Adenine mutagenesis could occur at several different steps including during synthesis or potential modification of the template RNA, minus-strand cDNA synthesis or plus-strand cDNA synthesis. During the analysis of BPP DGR mutagenic homing, where the donor TR was tagged with a self-splicing group I intron, TR-RNA intermediates were observed without adenine-specific mutagenesis suggesting that sequence diversification does not result from site-specific mutation or modification of the RNA template. To capture the DNA product as it integrates into the host chromosome, a phage recipient which only had the IMH element of VR was targeted for mutagenic homing. While cDNA integration at the 3' end occurred, the homing product appeared to be locked in this intermediate step as no 5' cDNA integration was detected. Sequence analysis of these intermediate step cDNA products revealed they contained adenine-specific mutagenesis with patterns essentially identical to those observed in DGR homing products. Furthermore, adenine-specific mutagenesis was observed in cDNA products synthesized in the absence of the target DNA, indicating that adenine mutagenesis likely occurs during minus-strand cDNA synthesis.

The above observations and the fact that *RT* is the only protein encoding gene strictly conserved to all DGRs suggests it is responsible for adenine-specific mutagenesis. The Atd protein forms an hourglass-shaped homopentameric structure with many DGRs containing analogous genes and single mutations at codons encoding conserved amino acid residues only effects the efficiency of mutagenic homing but has no effect on adenine-specific mutagenesis [20]. DGR RTs vary in size but all have a conserved central core that includes common structural motifs found in most other RTs with variable amino acid sequences at their N- and C-terminus. However, DGR RTs lack domains associated with RNaseH activity, like in retroviruses, or the endonuclease activity, similar to certain mobile group II introns and LINEs. Within the core regions, DGR RTs appear to have some interesting sequence features. The most prominent is located within the finger 4 region, containing a **G***L***PIG***NLT***SQ** (bold, highly conserved amino acids; italic, less conserved) motif, with G1, I4 and Q10 conserved only in DGR RTs. This motif corresponds to the nucleotide binding pocket of the HIV-1 RT that positions incoming dNTP substrates, influencing specificity and error-prone polymerization. The HIV-1 RT residue Gln151, which corresponds to the isoleucine residue (I4) of the DGR motif, was found to confer AZT (nucleoside analog) resistance to the HIV-1 RT. It is possible that this unique, conserved motif of DGR RTs is partly responsible for adenine-specific mutagenesis.

There are currently two hypotheses for the mechanism of adenine-specific mutagenesis. Adenine-specific mutagenesis may be a result of error-prone reverse transcription by DGR RTs that inserts random, standard deoxyribonucleotides opposite to the adenine residues in the template RNA. RTs are known to be error-prone during

reverse transcription and DGR RTs may have acquired a structural variation resulting in a lack of fidelity during the recognition/incorporation of adenines. Alternatively, DGR RTs cDNA synthesis could incorporate dUTP when reverse transcribing adenine residues in the template RNA. dUTP residues in the cDNA products would then be recognized by host-encoded uracil DNA glycosylases (UDGs) and excised, leaving abasic sites in the minus-strand cDNA. During plus-strand cDNA synthesis, which could be catalyzed by a DGR RT or a host-encoded DNA polymerase, random nucleotides are then incorporated opposite to these sites.

The contribution of retroelements to host fitness

One of the most fascinating aspects of DGRs is their ability to enhance host fitness through the accelerated evolution of TP ligand binding domains [14, 15]. In contrast, retroelements as a class are generally considered to be detrimental to host fitness. TEs found in eukaryotes comprise between 22% to 50% of their host genomes as compared to the 3% occupied by protein encoding sequences [7]. Retroelements comprise more than 90% of human TEs, with L1 found at ~500,000 genomic copies, and their mobility throughout the genome can have a number of effects ranging from insertions that result in gene silencing, they can serve as a platform for homologous recombination, DNA mediated transduction events between gnomic loci, generation of pseudogenes, acting as alternative promoters, acting as a transcriptional silencer or enhancer, their insertion can result in novel exon splicing, as well as many other functions [7, 8, 22]. The mobility of retroelements has resulted in human diseases such

as hemophilia A and elliptocytosis which have been well documented [7, 23] however, retroelement mobility in humans has been attributed to the beneficial tissue specific expression of amylase and recent evidence suggests they may function in beneficial gene regulation [8], making their relationship with host fitness unclear [24]. Group II introns are found within essential and non-essential genes in a quarter of sequenced bacterial gnomes as well as in eukaryotes. While it has been proposed that group II introns are the evolutionary ancestor of extant retroelements, the precise role of these elements, beyond proliferation and their effect on genome evolution, is currently unknown [9, 10]. Retrons are another group of retroelements where the exact nature of their relationship with the host is unknown, yet is through to be fairly innocuous. While the function of retrons beyond the production of msDNA is unknown, the deletion of *ret* and presumably all retron activity has no detectable effect on host fitness [11]. In contrast to most retroelements, the *Bb* BPP DGR can generate a vast repertoire of tail fiber receptor-host ligand interactions allowing tropism switching and a rapid adaptation to the dynamic *Bb* cell surface. This demonstrates that the *Bb* DGR, and perhaps all DGRs, has traded retrotransposition based mobility in order to generate diversity within proteins which function in ligand-receptor interactions, a powerful selective advantage for the host [14, 15, 17].

The breadth of DGRs within the tree of life

DGRs are widely distributed. To date the *Bb* bacteriophage BPP has served as the paradigm for DGRs in both mechanistic and functional studies, yet the ability to

iteratively diversify protein ligand binding domains should be of broad utility to a number of organisms [14, 15, 19]. Bioinformatic analysis, by us and others, of nucleotide databases was performed using custom made algorithms (Figure 3) to identify putative RTs through the identification of DGR specific domains (described above). The flanking nucleotide sequences were then searched to identify small repeat sequences which vary from each other at positions corresponding to adenines. These sequences are curated for small proteins, corresponding to accessory proteins, as well as DGR components e.g. IMH, DNA stem loops, etc.

These analyses revealed several key observations about the distribution of DGRs within the tree of life. DGRs are found distributed broadly across the bacterial domain as chromosomal, phage, and plasmid elements [12, 14]. To date over 244 species, representing at least 20 phyla have been identified in organisms that span the gamut of habitats and niches; from bacteriophage, to free-living nitrogen fixing marine bacteria, to human gut commensals, to human and plant pathogens (M. Gingery personal communication). Although DGRs are found broadly, they are not always found deeply as there appears to be an enrichment of elements in the three phyla *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* [12, 25]. However, it is not clear if this enrichment is significant or a result of sequencing bias as these phyla have the greatest number of deposited genomes for analysis. The DGRs found in sequenced genomes also show a variety of architectures but retain components known to be necessary for mutagenic homing.

The distribution of DGR has also been investigated in metagenomic datasets which identified a large number of DGR associated genes. An analysis to identify bacteriophage that infect gut commensal bacteria was performed using high-throughput sequencing of the intestinal virome of healthy human individuals and identified DGRs within 11 out of the 12 subjects studied [26]. Viral particles were isolated from stool samples and sequence analysis of phage genomes identified 51 highly variable regions demonstrating 96% variance at the amino acid level indicating almost every phage genome contained unique sequences. Interestingly, of these 51 hypervariable regions, 36 contained TR/VR pairs of which 29 contained an adjacent RT, and these regions were found to share homology with the BPP DGR. This and contextual data suggests many of these 29 regions represent bacteriophage with likely the same lifestyle as *Bb* BPP. Furthermore, 6 hypervariable regions were found within genes predicted to have homology to immunoglobulin (Ig)-superfamily β-sandwitch domains [26]. While genes encoding Ig domains have been previously reported in phage [27], this was the first report of a DGR diversified VR within an Ig domains, and what was most striking was the VR was found in the middle of the gene [26]. Due to the increased distance from TR to VR as well as the apparent lack of DNA stemloop/cruciform structures, it is unclear if Ig domain containing TPs are diversified by a similar mechanism as BPP. Structural studies of *Bb* Mtd revealed the CLec fold was a common solution among DGRs to balance the ability to display diversity within a rigid backbone [28] and this work adds to the paradigm for protein solutions for displaying diversity as these elements would represent the first instance of Ig domain diversification through a RT-based mechanism. This analysis demonstrated that DGRs are common in the genomes of bacteriophage

found in the lower gastrointestinal tract of humans, these elements have coopted multiple protein fold for display, and these phage display high levels of mutagenic homing [26].

Analysis of metagenomic datasets also identified the first DGRs outside the bacterial domain. Single cell sequencing of samples isolated from a mine in South Dakota identified several DGRs within the genome of *Nanoarchaeum equitans* (Blair Paul personal communication). While analysis is ongoing these elements contain all known genes and factors necessary for mutagenic homing. Interestingly, the *N. equitans* DGR can encode multiple TPs which display 3' VRs, similar to *mtd*, and TPs which display mid-gene DGRs, similar to phage of the human gut. Multiple, independent studies have confirmed that DGRs are widely distributed throughout the bacterial domain, with TPs predicted to have diverse functions, and have been described within two domains of life suggesting these elements are of broad utility.

DGRs are distributed throughout bacterial populations using a variety of mechanisms. DGRs appear to have been distributed throughout the bacteria domain by both vertical inheritance and horizontal gene transfer. Vertical inheritance of DGRs is inferred when a group of DGRs with similar characteristics to each other are found within a group of closely related organisms and analysis typically involves the key DGR protein, the RT, having amino acid similarity to other reverse transcriptases (DGR or otherwise) in the host genome [12]. Additionally, vertical inheritance is inferred when the DGR is found to have a G+C content within 3% of its host genome. In contrast, analysis of several DGRs suggests they have been acquired by horizontal gene transfer.

Elements have been identified on phage and plasmids which have a high variance in G+C content from their host. Interestingly, a DGR has been identified in a *Vibrio* phage where the element appears to have originated in the BPP phage and recombined into the kappa prophage [12]. DGRs have also been found as part of genomic islands which encode transposase- or integrase-like sequences as well as within Integrative and Conjugative Elements (ICE) [29], all well described means of gene transfer.

DGRs have been identified widely within the bacteria domain and it appears they are just as likely to be inherited vertically as by horizontal gene transfer. It is unclear which species, phage or bacteria, was the nucleating point(s) for the distribution of DGRs however it is clear that, while some elements have coevolved with their host genomes over evolutionary time, others have been recently acquired but it is likely that all of these elements have been retained because of the selective advantage they offer their hosts.

A DGR was identified within *Legionella pneumophila*

Legionella pneumophila. *In silico* analysis of nucleotide databases (see above) identified a putative DGR within the genome of one of five sequenced *Legionella pneumophila* (*Lp*) strains called Corby and my work detailing this element will be described in greater detail in forthcoming chapters.

While the *Legionella* genus is made up of more than 40 species, *Lp* is responsible for >90% of disease [30] and while *Lp* is composed of 15 serogroups, serogroup I (SgI) accounts for ≥84% of cases of disease. Interestingly, SgI was

previously thought to be primarily composed of clinical isolates however recent work has shown that both clinical and environmental isolates are well represented with both isolates being able to cause disease in humans [31]. *Lp* is an opportunistic human pathogen and etiological agent of Legionnaires' disease which is an uncommon form of pneumonia with no distinguishing clinical features identifying it from other types of pneumonia [32]. It is found ubiquitously in soil and freshwater environments where it survives by parasitizing protists, forming biofilms, and after accidental inhalation by humans, replicating within alveolar macrophages or monocytes [33-35]. *Lp* can be actively endocytosed or can induce their uptake through using an unknown mechanism or through an alternative mechanism called "coiling" phagocytosis [36]. After uptake, *Lp* actively prevents the typical phagosome maturation through the action of a type 4 secretion system (T4SS). This system subverts host cellular processes in order to mature the endosome into a *Legionella* containing vacuole (LCV) which associates with mitochondrial and endoplasmic reticulum derived vesicles [34]. The bacteria thrive and replicate within the LCV until they sense depleted nutrients, whereas they lyse the cell, seek a new host, and the infection process begins anew [37]. The ability of *Lp* to survive and replicate within aquatic environments is explained by its biphasic lifestyle. *Lp* transitions between distinct phenotypic states, from a non-mobile intra-cellular replicative state to a mobile transmissive state which is stress resistant, making them well suited to survive in man-made chlorinated environments [38]. Interestingly, *Lp* can infect most single celled organisms as well as mammalian cells and their lifecycles are morphologically indistinguishable suggesting that conserved mechanisms are targeted to facilitate survival in both species [34, 39]. While *Lp* maintains a core set of genes and

systems necessary for pathogenesis as well as environmental survival, the degree of genomic plasticity observed between strains is striking especially since the effect this plasticity has on virulence is unknown.

Legionella pneumophila genomes. While there are over 15 serogroups of *Lp*, most genomic and genetic analysis has been performed on members of SgI due to their ability to cause disease in humans. To date there have been seven sequenced *Lp* SgI isolates whose circular genomes are ~3.5 megabases (Mb), with an average G+C content of 38%[21]. The *Lp* genomes are larger than many intracellular pathogens and it has been suggested that these larger genomes reflect the need for genes which facilitate its adaptation to multiple environments [39]. Only two strains, Pairs and Lens, maintain large plasmids however all strains maintain episomal elements. An analysis of four *Lp* genomes identified a high level of genetic synteny with only one large inversion. Additionally, the core genome for *Lp* strains was found to comprise 80% of genes while unique, strain specific genes comprise between 7-11% of the genome (Figure 4). While this degree of strain specific genes is high for intracellular pathogens, it is comparable to other free living proteobacteria [39]. As part of the core genome, all *Lp* strains encode at least two secretion systems which are necessary to invade host cells. The T4SS encoded by the *dot* and *icm* genes secretes >300 effectors which usurp host cellular targets, such as GTPases, and modulate cellular processes, such as protein secretion or apoptotic signaling pathways and mutations in this secretion system result in *Lp* which cannot prevent the maturation of phagosomes or consequently prevent their fusion with lysosomes [40]. A second feature, found in the core genome, is a type 2 secretion system (T2SS) which has been shown as key for the secretion of several

enzymes and whose deletion results in cells unable to replicate under stressful conditions [41]. While *Lp* encodes a core set of proteins necessary for its survival, it is interesting to note that each strain has a genetic uniqueness to it and tempting to speculate that this individuality contributes to the wide degree of clinical manifestations seen between *Lp* strains [30, 42].

Legionella pneumophila strain Corby. *Lp* strain Corby is a SgI clinical isolate that has been described as hyper-virulent. When animals are infected with various strains of *Lp* via an aerosol challenge, only strain Corby replicated to high numbers in the lungs and caused pyrexia which was followed by death within days. Furthermore, it was observed that Corby cells would replicate to high numbers within phagocytic immune cells, while other strains were mostly observed extracellularly. A possible explanation for this difference is that Corby has an increased capability to invade host cells or is better able to resist degradation by the host however, no molecular mechanism has been presented to explain this apparent increase in virulence [42]. While, each *Lp* strain has a unique set of genes it is interesting to note that Corby contains protein secretion systems and mobile genetic islands not found in other sequenced strains [29, 43].

Protein secretion in Gram negative bacteria

DGR mutagenic homing does not function on VRs in isolation; they do not diversify solely for the sake of generating nucleotide variability. The VR is found within protein domains which are thought to balance the display of peptide variability with a

rigid scaffold so that DGR-generated diversity is presented in a contextually meaningful way [17, 28]. It is likely that selection is acting upon diversified TPs and knowledge of subcellular locality of diversified TPs may lead to a better understanding of putative functions. To this end, we will briefly discuss protein section in gram negative (-) bacteria.

Major protein secretion systems. Gram- bacteria are compartmentalized by lipid bilayers and this compartmentalization is necessary for the function of most cellular processes. Proteins and factors must be transported between these compartments, from the interior of the cell to the exterior and vice versa. Several general protein secretion systems have been described which include type 1-6 secretion systems as well as dedicated secretion systems such as those responsible for transporting β-barrel proteins, lipoproteins, or lipopolysaccharides (LPS) [44, 45]. The general secretory system (SEC) translocates unfolded proteins across lipid bilayers and is conserved across the three domains of life. In gram- bacteria, this system is used to translocate proteins across or insert proteins into the inner membrane [46]. SEC transport can occur during or after protein synthesis and in either case requires binding of the polypeptide sequence by chaperone molecules. In post-translational transport, these chaperones can be either SecA or SecB [47]. The bound peptide is delivered to the complex of proteins, SecYEG, which form a channel across the bacterial inner membrane. In some instances, the SecYEG pore has been shown to interact with a second complex SecDFYajC however, since translocation can occur only via SecYEG, it is unclear what role this second complex plays [46]. ATP hydrolysis by SecA is required for threading the protein through the channel and, after translocation through

the pore, the protein is available for ligand interaction or additional modification by host systems [47].

An alternative protein secretion system that is present in many bacteria, Archaea, some mitochondria, and thylakoid membranes is the twin-arginine translocation (TAT) system. The TAT system has the unique ability to transport folded proteins, or protein complexes, across lipid bilayers all while maintaining selective permeability [48]. TAT translocation transports proteins post-translationally and in *Escherichia coli* requires two trans-membrane containing proteins, TatB and TatC, which play a role in substrate recognition. A third transmembrane helix containing protein, TatA, is thought to be present within the bacterial inner membrane as monomers until substrate is bound by TatBC whereas they oligomerize into the major pore forming unit [44, 48]. This dichotomy of steady state monomerization versus triggered pore formation of TatA could explain how cells are able to maintain membrane selectivity for small ions/molecules and still translocate proteins up to ~70A in diameter. The TAT system is not thought to be required for life and thus is found distributed throughout the three domains, and has undergone convergence, e.g. some organisms do not encode for *tatB*, instead TatA in conjunction with TatC, fulfills a dual purpose [48].

Protein signal sequences are recognized by host secretion systems. While these are only two of the many bacterial secretion systems, they are instrumental in the translocation of proteins across the inner membranes of gram negative bacteria. Proteins transported by these systems are synthesized with a signal sequence/peptide that is necessary for their recognition [45]. SEC secreted proteins contain N-terminal

signal sequences that are typically 20 to 30 amino acids, although can be longer, and are divided into three distinct domains [47]. These signal peptides begin with one to eight basic, positively charged residues then a domain comprised of up to 16 hydrophobic amino acids followed by a highly variable stretch of less polar residues which typically contain a peptidase cleavage site. Preproteins or immature protein, proteins whose signal sequences have not been cleaved off, are recognized via their signal sequence by a variety of chaperones and delivered to the SecYEG pore [47].

Similarly to SEC substrates, TAT translocated proteins have a tripartite signal sequences with positive/hydrophobic/polar domains. However, since the TAT system only transports folded proteins or protein complexes, there are distinguishing features which discriminate the recognition of substrates by this system. The positively charged domains of TAT substrates contain a peptide consensus motif of SRRxFLK, where “x” is any polar amino acid. This motif is structured around the two arginine residues that are required in plant signal peptides but are variable in bacterial signal peptides. However, bacterial TAT substrates must contain the first arginine with mutations in the second arginine typically affecting translocation rate [49, 50]. Additionally, the hydrophobic domains of TAT substrates are less hydrophobic than SEC substrates and changes in this domain result in recognition by the wrong secretion system. Finally, TAT substrates contain basic residues in their polar domain which are not common in SEC substrates [48]. Unlike the SEC system, there does not appear to be dedicated chaperones that recognize and deliver preproteins to the TAT translocon instead protein signal sequences are recognized directly by TatBC. However, there are reports of chaperones

which are important in proofreading to ensure native folding and it is unclear if these proteins play a role in substrate delivery to the TAT system [48].

Trafficking of lipoprotein by gram negative bacteria. Lipoproteins are a class of proteins with lipid moieties conjugated onto conserved residues found in the inner- and outer-membrane of gram negative bacteria and have been implicated in a wide variety of cellular processes [45]. They are transported across the inner membrane by either the SEC or TAT system by recognition of an N-terminal signal peptide and, once translocated, these signal sequences are modified by cellular machinery found at the periplasmic/inner membrane interface [45, 48]. Lipoprotein signal peptides contain a consensus motif, called the lipobox, within the polar domain. In *E. coli*, lipobox motifs are typically comprised of amino acids L-(A/S)-(G/A)-C and, while the first three residues can be variable, the fourth is required to be a cysteine [45]. In gram- bacteria once lipoproteins have been translocated across the inner membrane, three highly conserved enzymes are responsible for cleavage of the signal peptide and modification of the conserved cysteine. The first, *lgt*, encodes a phosphatidylglycerol transferase and is responsible for the formation of a thioester linkage between the conserved cysteine of the lipobox and a diacylglycerol moiety. This initial modification is followed by cleavage of the signal peptide between the third and fourth lipobox residue by the lipoprotein specific signal peptidase II, LspA. This leaves the conserved cysteine to be further modified by the conjugation of a fatty acid moiety by action of the phospholipid transacylase, Lnt [45]. These cleaved and modified lipoproteins are considered to be mature proteins. While there have been reports of mature lipoproteins where the signal sequence has not been removed by LspA, the modification of the lipobox conserved

cysteine is thought to be necessary for the retention of lipoproteins within lipid bilayers [47]. Systems for lipoprotein modification are thought to be necessary for life and are highly conserved across bacterial species with homologues being described in gram+ bacteria [45].

Types of modifications found on bacterial proteins. Post-translational modification of proteins, including lipoproteins, is a widely studied field due to the importance these modifications have on protein function and, while this work has mainly focused on eukaryotic systems, several examples of protein modification in prokaryotic systems have been described [51]. Prokaryotes are able to modify their protein at a variety of residues with a wide variety of modifications which includes: acetylation, carboxylation, deamidation, glycosylation, lipidation, methylation, phosphorylation, proteolysis, pupylation, ribosylation as well as others [51]. The breadth of bacterial PTM and their effect on host processes are constantly expanding [52].

Delivery of lipoproteins to the outer membrane requires LOL. The observation that lipoproteins were found in both the gram negative outer- and inner-membranes lead to the discovery of a dedicated transport system which ferries lipoproteins across the aqueous periplasmic space. This system, called the localization of lipoprotein (Lol), is comprised of five proteins, with an ATP binding cassette complex being formed by LolCDE at the inner membrane. This complex binds outer membrane destined, matured lipoproteins and ATP hydrolysis leads to a conformational change which results in the transfer of the lipoprotein to the periplasmic transport protein LolA [53]. The LolA-lipoprotein complex transverses the periplasmic space and, through an

affinity based interaction, transfers the lipoprotein to the outer membrane protein LolB which inserts the lipoprotein into the bacterial outer membrane through an unknown process [45, 53].

Several studies have investigated the molecular mechanisms by which lipoproteins destined for the bacterial inner- or outer-membrane are recognized and subsequently transported by the Lol system. The N-terminal amino acids of lipoproteins have been systematically investigated to determine their contribution to protein transport by the Lol system. Lipoproteins have a modified conserved cysteine which, when the signal sequence is cleaved off, becomes the first or, using the naming convention, the +1 residue. It has been demonstrated that the chemical characteristics of amino acids in the next two positions or +2/+3 influence to which membrane proteins are delivered to [45]. The systematic substitution of amino acids into these two positions has been performed and the results suggest a number of rules that are still being debated however, the nuances are beyond this report. The characteristics of amino acids in the +2 play a major role in the retention of lipoproteins in the inner membrane. Specifically, lipoproteins with aspartic acid at the +2 residue are typically found in the inner membrane [54]. Furthermore, when outer membrane lipoproteins have their +2 residue replaced with either aspartic acid, phenylalanine, tryptophan, threonine, glycine, or proline they are retained in the inner membrane [45]. Conversely, mature lipoproteins with a serine in the +2 position are typically found in the outer membrane. Interestingly, the characteristics of +3 residues was found to influence membrane localization, as the substitution of serine into the +3 of inner membrane proteins resulted in their outer membrane localization, except for when aspartic acid was at the +2 position. The

placement of histidine or lysine at the +3 position with aspartic acid in the +2 position resulted in partial retention of lipoproteins in the inner membrane, suggesting the characteristics of both residues in the +2/+3 position interact with Lol machinery to influence trafficking [45].

It has been suggested that all lipoproteins, regardless of membrane destination, are recognized by LolCDE through their modified conserved cysteine. Additionally, experiments where the negative charge of the +2 aspartic acid was chemically neutralized or where the +2 serine was chemically oxidized resulted in altered membrane localization. This suggests LolCDE interacts will all lipoproteins and the physical distance from the conserved cysteine to the negative charge of the +2 residue, in outer membrane destined proteins, might be responsible for initiating ATP hydrolysis by LolCED, followed by subsequent transfer of the lipoprotein to LolA [55].

Characterization of a DGR within Legionella pneumophila strain Corby.

Through our analysis of deposited nucleotide databases a retroelement was identified in *Lp* strain Corby which encodes all components that are characteristically shared amongst DGRs.

Beginning in chapter 2 we will describe our analysis of the Corby DGR and demonstrate this putative element is capable of mutagenic homing and diversification of its target gene, *ldtA*. Our analyses revealed the DNA, RNA, and protein requirements for mutagenic homing in the *Lp* and BPP DGRs are analogous, implying mechanistic

conservation. Observations of mutagenic homing during the biphasic lifecycle of Corby lead us to hypothesize *Lp* DGRs may be regulated. We took advantage of the extensively studied regulatory cascades that control *Lp* phenotypic progression in order to analyze how Corby has integrated regulation of its DGR into global signal cascades.

In chapter 3 we will discuss the distribution of DGRs in the *Legionella* genus. By screening a *Lp* library of clinical isolates we identified elements that are highly homologous to the Corby DGR but have co-opted different carrier sequences to display related variable domains. Additionally, we have described a DGR in the related species *L. tunisiensis* and will discuss its relationship to elements found in *Lp*. Finally, an examination of TPs found in diverse gram-negative bacteria suggests that lipoprotein anchoring and surface display of DGR-diversified protein repertoires may be a common theme in gram-negative bacteria.

Then in chapter 4 we will discuss the target of DGR mutagenic homing the Corby TP, LdtA. We identified several conserved domains, an N-terminal bipartite signal peptide containing non-canonical secretion motifs as well as a C-terminal CLec domain. LdtA was demonstrated to be a surface displayed lipoprotein whose variable domain is exposed to the extra-cellular milieu, available for ligand interaction. With the surface display of lipoproteins being rare we wanted to identify necessary systems and motifs to understand why certain proteins are trafficked to the cell surface. Surface display of LdtA fuses the TAT and LOL pathway in an unusual manner that is currently being investigated. We then finish with the implications of LdtA as a TAT-lipoprotein and its secretion in related systems.

Finally, in chapter 5 we will briefly comment on the results of all chapters, discuss current as well as future experiments, and finish with the importance of this work in regards to understanding the biology of DGRs.

Figure Legends:

Figure 1. Relationship of DGRs to other retroelements.

A. Major retroelement groups are shown, along with their identification in bacteria, bacteria-derived organelles, and the eukaryotic nucleus. LTR elements include *LTRs & retroviruses, hepadnaviruses*; non-LTR elements are *TERTs (telomerase RTs), PLEs (Penelope-like elements), non-LTRs, group II introns, DGRs, and retrons*. **B**. A phylogenetic tree depicts the relationship of DGRs to other prokaryotic retroelements. The tree was constructed with representative RT sequences from seven subclasses of group II introns (B-F, ML, CL), retroplasmids, non-LTR elements, and retrons. These RTs were selected because they are the most closely related to DGRs, and provide enough alignable characters to allow a minimal degree of resolution (120 aa), whereas adding more RTs reduced the number of characters and prevented resolution. The tree was constructed with the RAxML algorithm used with the RtREV model, and rooted with retrons, and numbers indicate either bootstrap support (upper numbers) or posterior probabilities from Bayesian analyses (lower numbers). The lengths of the triangles represent the number of substitutions per site for the longest branch within each clade. Widths do not represent the number of sequences in the group. Black dots indicate the three supported nodes that place DGRs internal to group II introns.

Figure 2. Model for DGR mutagenic homing in *Bordetella bronchiseptica* phage BPP.

The current model for DGR mutagenic homing in the *Bb* phage BPP has been proposed to operate by a TPRT mechanism. Through the action of an unidentified factor TR-RNA is transcribed and a single-stranded break occurs within the VR IMH (purple) exposing a 3'-OH. While the exact mechanism is currently under investigation, VR target recognition requires stem loop/cruciform structure (purple stems) and the exposed 3'-OH operate to prime DGR-RT dependent (red oval) cDNA synthesis using the TR-RNA transcript (blue) as a template. During reverse transcription TR adenines are randomly changed to any of the four deoxyribonucleotides (N) in the resulting cDNA (dashed line). Integration of the cDNA into the chromosome occurs at the 3' break in the IMH and is dependent upon 5' homology and results in the displacement of the parental strand with a diversified VR. *Avd* encodes a protein necessary for mutagenic homing that is hypothesized to interact with nucleic acids. The VR containing Mtd (colored circles on phage) is integral for phage attachment to *Bb* cell surface ligands and its diversification by the DGR expands its repertoire of ligands.

Figure 3. Flow chart demonstrating the identification of DGRs within nucleotide sequence databases.

(1) Protein databases were searched with Bordetella phage BPP-1 Brt sequence, then iteratively with BRT homologs for outlier DGR-RT-like proteins (BLAST or PSI-BLAST). Output was filtered with a DGR-RT-specific motif to screen out other RT types. (2) Regions ±10 kb around a putative *DGR-RT* gene were searched for direct repeats, noting that one repeat (TR) should be very near, or within the N-terminus of the *RT*

gene, and the other (VR) should be at the C-terminus of a nearby ORF. (3) VR was aligned to TR to determine that differences are mainly at TR adenine residues. (4) Regions near *RT* were searched for an *Avd* gene. (Most, but not all, DGRs will have an *Avd* gene (and/or an HRDC domain-containing gene) present next to *RT*). (5) To find remote target genes in distant regions of a genome, TR nucleotide sequence were BLASTed against the host organism genome sequence to find remote VRs, and thus remote target genes.

Figure 4. Comparison of *Lp* genomes

A Venn diagram comparing four *Lp* SgI genomes. The core genome encoded by all four strains is indicated by overlapping circles. Strain specific genes with percentage of the total genes is encoded by these genes are indicated. The table lists strains along with genome size and total number of predicted open reading frames.

Figure 1

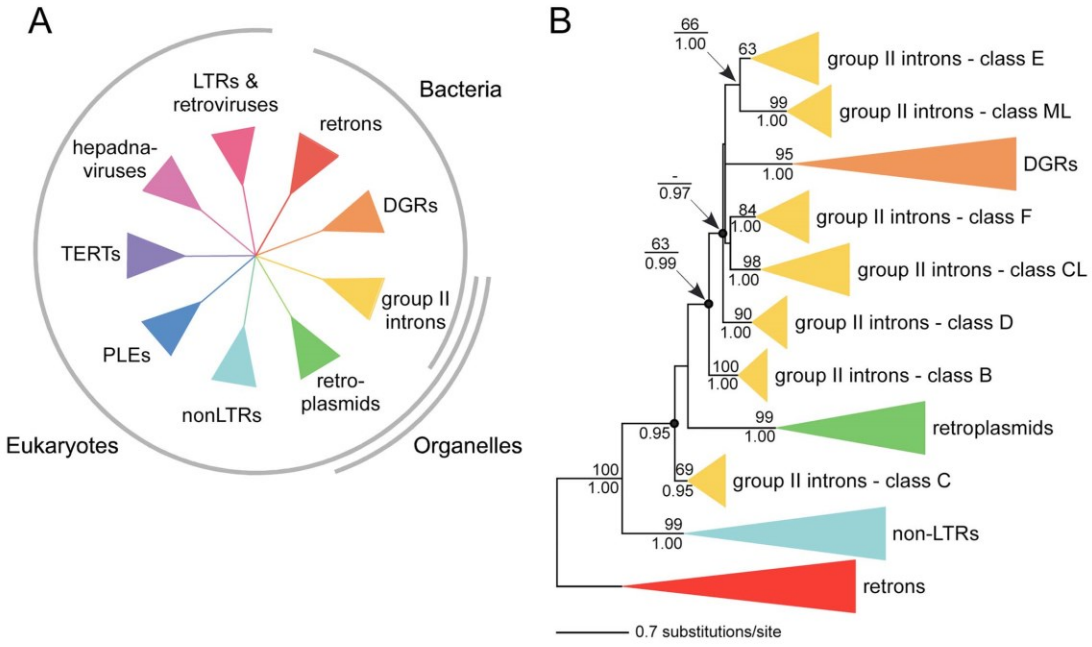

A
LTRs &
retroviruses
Bacteria
retrons
hepadna-
viruses
DGRs
TERTs
group II
introns
PLEs
retro-
plasmids
nonLTRs
Eukaryotes
Organelles
B
66
1.00
63
group II introns - class E
99
1.00
group II introns - class ML
95
1.00
DGRs
-
0.97
84
1.00
group II introns - class F
63
0.99
98
1.00
group II introns - class CL
90
1.00
group II introns - class D
100
1.00
group II introns - class B
99
1.00
retroplasmids
0.95
100
1.00
69
0.95
group II introns - class C
99
1.00
non-LTRs
retrons
0.7 substitutions/site

Figure 2

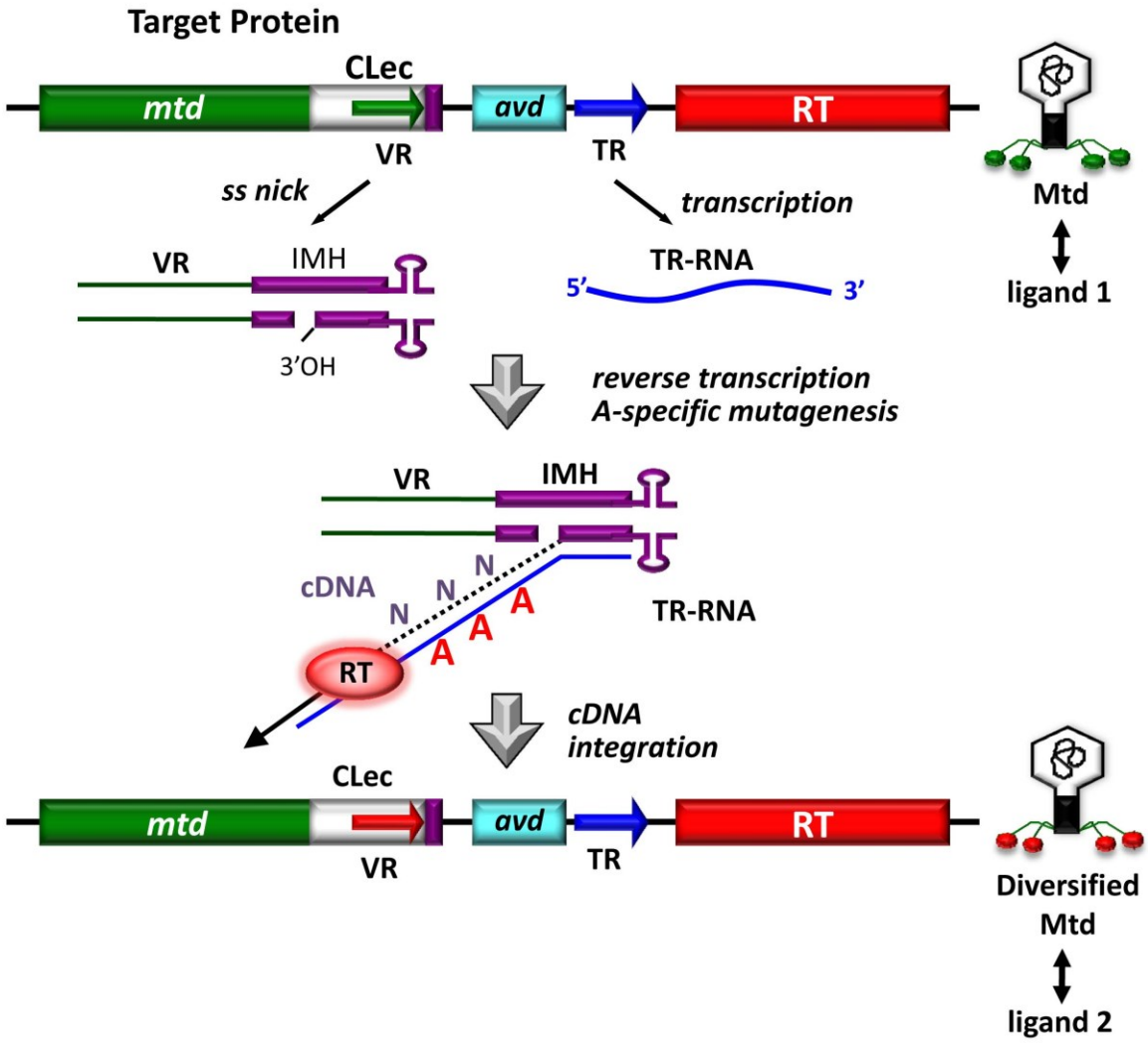

Target Protein
CLec
mtd
avd
RT
VR
TR
Mtd
ligand 1
ss nick
transcription
VR
IMH
3'OH
TR-RNA
5'
3'
reverse transcription
A-specific mutagenesis
VR
IMH
cDNA
N
N
N
A
A
A
TR-RNA
RT
cDNA
integration
CLec
mtd
avd
RT
VR
TR
Diversified
Mtd
ligand 2

Figure 3

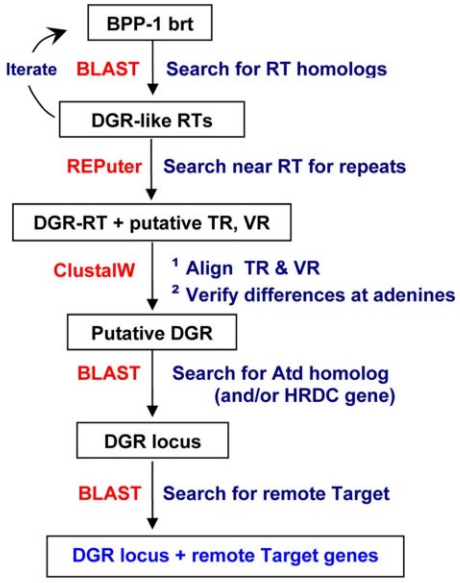

BPP-1 brt
Iterate
BLAST
Search for RT homologs
DGR-like RTs
REPuter
Search near RT for repeats
DGR-RT + putative TR, VR
ClustalW
1 Align TR & VR
2 Verify differences at adenines
Putative DGR
BLAST
Search for Atd homolog
(and/or HRDC gene)
DGR locus
BLAST
Search for remote Target
DGR locus + remote Target genes

Figure 4

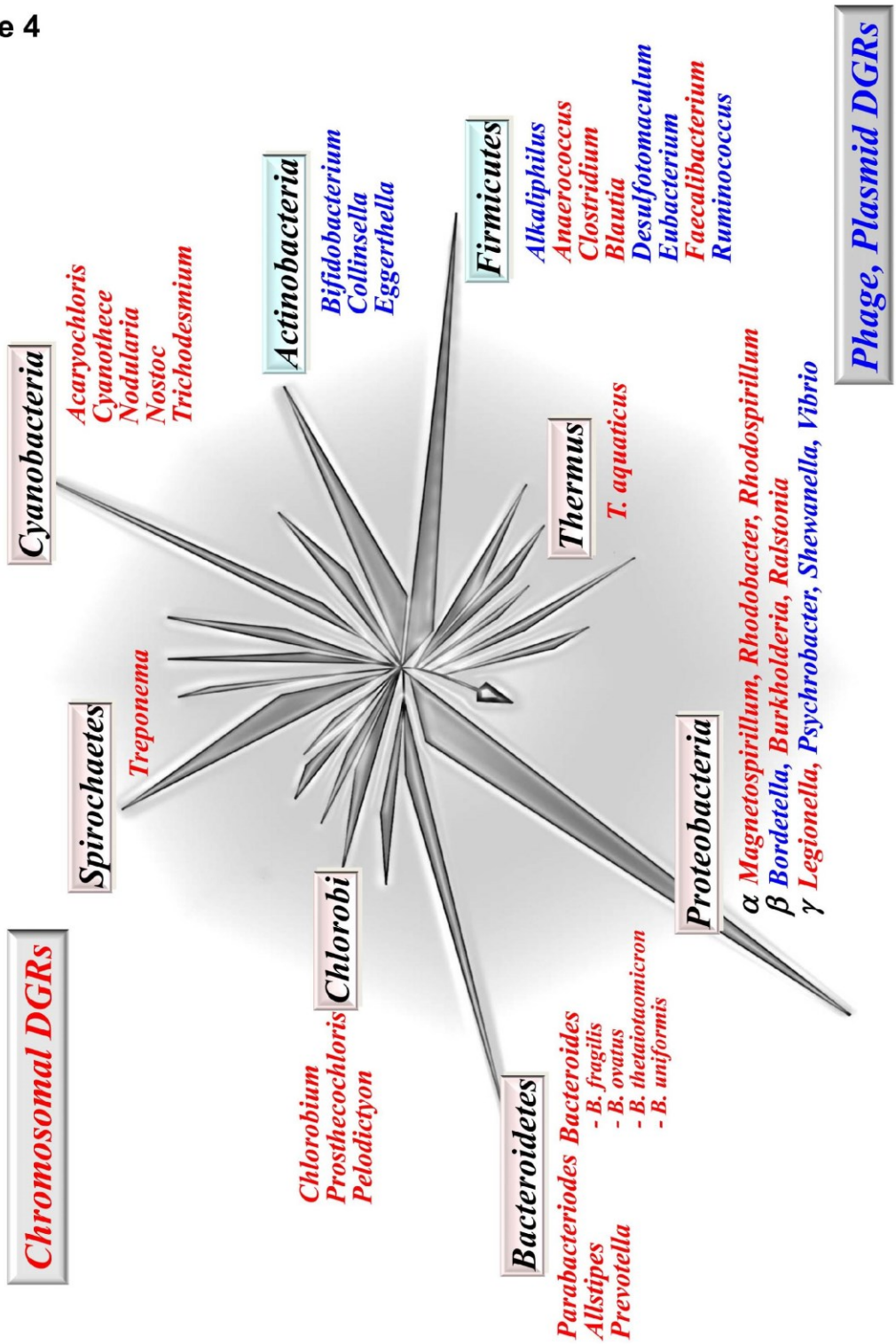
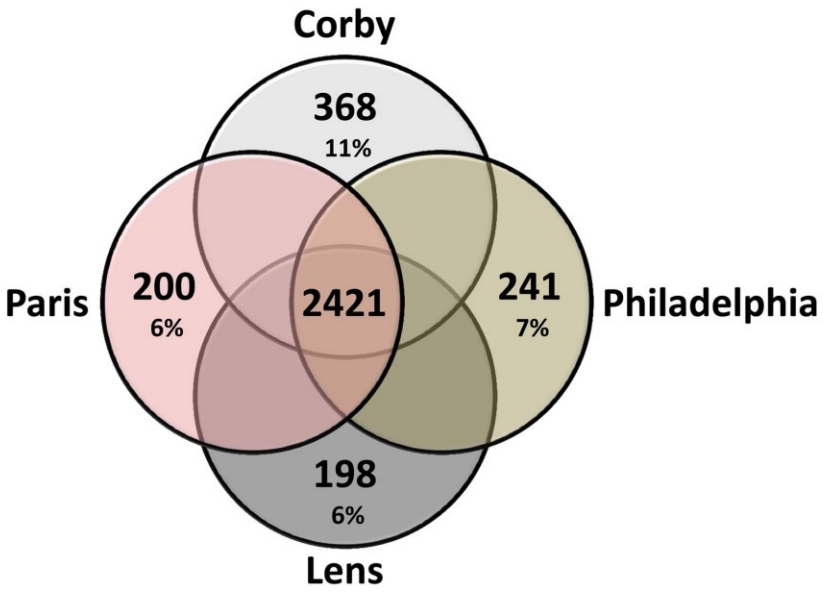

Chromosomal DGRs
Cyanobacteria
Acaryochloris
Cyanothece
Nodularia
Nostoc
Trichodesmium
Spirochaetes
Treponema
Actinobacteria
Bifidobacterium
Collinsella
Eggerthella
Firmicutes
Alkaliphilus
Anaerococcus
Clostridium
Blautia
Desulfotomaculum
Eubacterium
Faecalibacterium
Ruminococcus
Thermus
T. aquaticus
Chlorobi
Chlorobium
Prosthecochloris
Pelodictyon
Bacteroidetes
Parabacteriodes
Allstipes
Prevotella
Bacteroides
- B. fragilis
- B. ovatus
- B. thetaiotaomicron
- B. uniformis
Proteobacteria
α Magnetospirillum, Rhodobacter, Rhodospirillum
β Bordetella, Burkholderia, Ralstonia
γ Legionella, Psychrobacter, Shewanella, Vibrio
Phage, Plasmid DGRs

Figure 5



Strain:	Genome Size (Mb):	Gene:
Corby	3.58	3,257
Philadelphia	3.49	3,184
Lens	3.41	3,058
Paris	3.64	3,278

References

1. Gordo, I., L. Perfeito, and A. Sousa, *Fitness effects of mutations in bacteria.* J Mol Microbiol Biotechnol, 2011. **21**(1-2): p. 20-35.
2. Mozhayskiy, V. and I. Tagkopoulos, *Guided evolution of in silico microbial populations in complex environments accelerates evolutionary rates through a step-wise adaptation.* BMC Bioinformatics, 2012. **13 Suppl 10**: p. S10.
3. Ochman, H. and N.A. Moran, *Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis.* Science, 2001. **292**(5519): p. 1096-9.
4. Etcheverria, A.I. and N.L. Padola, *Shiga toxin-producing Escherichia coli: factors involved in virulence and cattle colonization.* Virulence, 2013. **4**(5): p. 366-72.
5. Roth, J.R. and D.I. Andersson, *Amplification-mutagenesis--how growth under selection contributes to the origin of genetic diversity and explains the phenomenon of adaptive mutation.* Res Microbiol, 2004. **155**(5): p. 342-51.
6. Casacuberta, E. and J. Gonzalez, *The impact of transposable elements in environmental adaptation.* Mol Ecol, 2013. **22**(6): p. 1503-17.
7. Gogvadze, E. and A. Buzdin, *Retroelements and their impact on genome evolution and functioning.* Cell Mol Life Sci, 2009. **66**(23): p. 3727-42.
8. Rebollo, R., M.T. Romanish, and D.L. Mager, *Transposable elements: an abundant and natural source of regulatory sequences for host genes.* Annu Rev Genet, 2012. **46**: p. 21-42.
9. Toro, N., J.I. Jimenez-Zurdo, and F.M. Garcia-Rodriguez, *Bacterial group II introns: not just splicing.* FEMS Microbiol Rev, 2007. **31**(3): p. 342-58.
10. Lambowitz, A.M. and S. Zimmerly, *Group II introns: mobile ribozymes that invade DNA.* Cold Spring Harb Perspect Biol, 2011. **3**(8): p. a003616.
11. Lampson, B.C., S. Inouye M Fau - Inouye, and S. Inouye, *Retrons, msDNA, and the bacterial genome.* (1424-859X (Electronic)).
12. Schillinger, T., et al., *Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF.* BMC Genomics, 2012. **13**: p. 430.
13. Ding, W., et al., *L1 elements, processed pseudogenes and retrogenes in mammalian genomes.* IUBMB Life, 2006. **58**(12): p. 677-85.
14. Medhekar, B. and J.F. Miller, *Diversity-generating retroelements.* Curr Opin Microbiol, 2007. **10**(4): p. 388-95.
15. Liu, M., et al., *Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.* Science, 2002. **295**(5562): p. 2091-4.
16. Guo, H., et al., *Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification.* Mol Cell, 2008. **31**(6): p. 813-23.
17. Doulatov, S., et al., *Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements.* Nature, 2004. **431**(7007): p. 476-81.
18. Miller, J.L., et al., *Selective ligand recognition by a diversity-generating retroelement variable protein.* PLoS Biol, 2008. **6**(6): p. e131.
19. Guo, H., et al., *Target site recognition by a diversity-generating retroelement.* PLoS Genet, 2011. **7**(12): p. e1002414.

20. Alayyoubi, M., et al., *Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase.* Structure, 2013. **21**(2): p. 266-76.
21. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2010. **38**(Database issue): p. D5-16.
22. Carreira, P.E., S.R. Richardson, and G.J. Faulkner, *L1 retrotransposons, cancer stem cells and oncogenesis.* FEBS J, 2014. **281**(1): p. 63-73.
23. Ostertag, E.M., et al., *SVA elements are nonautonomous retrotransposons that cause disease in humans.* Am J Hum Genet, 2003. **73**(6): p. 1444-51.
24. Meisler, M.H. and C.N. Ting, *The remarkable evolutionary history of the human amylase genes.* Crit Rev Oral Biol Med, 1993. **4**(3-4): p. 503-9.
25. Schillinger, T. and N. Zingler, *The low incidence of diversity-generating retroelements in sequenced genomes.* Mob Genet Elements, 2012. **2**(6): p. 287-291.
26. Minot, S., et al., *Hypervariable loci in the human gut virome.* Proc Natl Acad Sci U S A, 2012. **109**(10): p. 3962-6.
27. Fraser, J.S., et al., *Ig-like domains on bacteriophages: a tale of promiscuity and deceit.* J Mol Biol, 2006. **359**(2): p. 496-507.
28. McMahon, S.A., et al., *The C-type lectin fold as an evolutionary solution for massive sequence variation.* Nat Struct Mol Biol, 2005. **12**(10): p. 886-92.
29. Lautner, M., et al., *Regulation, integrase-dependent excision, and horizontal transfer of genomic islands in Legionella pneumophila.* J Bacteriol, 2013. **195**(7): p. 1583-97.
30. Fields, B.S., R.F. Benson, and R.E. Besser, *Legionella and Legionnaires' disease: 25 years of investigation.* Clin Microbiol Rev, 2002. **15**(3): p. 506-26.
31. Kozak-Muiznieks, N.A., et al., *Prevalence of sequence types among clinical and environmental isolates of Legionella pneumophila serogroup 1 in the United States from 1982 to 2012.* J Clin Microbiol, 2014. **52**(1): p. 201-11.
32. de Jong, B., et al., *Travel-associated Legionnaires' disease in Europe, 2010.* Euro Surveill, 2013. **18**(23).
33. Declerck, P., *Biofilms: the environmental playground of Legionella pneumophila.* Environ Microbiol, 2010. **12**(3): p. 557-66.
34. Isberg, R.R., T.J. O'Connor, and M. Heidtman, *The Legionella pneumophila replication vacuole: making a cosy niche inside host cells.* Nat Rev Microbiol, 2009. **7**(1): p. 13-24.
35. Schroeder, G.N., et al., *Legionella pneumophila strain 130b possesses a unique combination of type IV secretion systems and novel Dot/Icm secretion system effector proteins.* J Bacteriol, 2010. **192**(22): p. 6001-16.
36. Bozue, J.A. and W. Johnson, *Interaction of Legionella pneumophila with Acanthamoeba castellanii: uptake by coiling phagocytosis and inhibition of phagosome-lysosome fusion.* Infect Immun, 1996. **64**(2): p. 668-73.
37. Al-Quadan, T., C.T. Price, and Y. Abu Kwaik, *Exploitation of evolutionarily conserved amoeba and mammalian processes by Legionella.* Trends Microbiol, 2012. **20**(6): p. 299-306.

38. Dalebroux, Z.D. and M.S. Swanson, *ppGpp: magic beyond RNA polymerase.* Nat Rev Microbiol, 2012. **10**(3): p. 203-12.
39. Gomez-Valero, L., C. Rusniok, and C. Buchrieser, *Legionella pneumophila: population genetics, phylogeny and genomics.* Infect Genet Evol, 2009. **9**(5): p. 727-39.
40. Shin, S., *Innate Immunity to Intracellular Pathogens: Lessons Learned from Legionella pneumophila.* Adv Appl Microbiol, 2012. **79**: p. 43-71.
41. Cianciotto, N.P., *Many substrates and functions of type II secretion: lessons learned from Legionella pneumophila.* Future Microbiol, 2009. **4**(7): p. 797-805.
42. Jepras, R.I., R.B. Fitzgeorge, and A. Baskerville, *A comparison of virulence of two strains of Legionella pneumophila based on experimental aerosol infection of guinea-pigs.* J Hyg (Lond), 1985. **95**(1): p. 29-38.
43. Glockner, G., et al., *Identification and characterization of a new conjugation/type IVA secretion system (trb/tra) of Legionella pneumophila Corby localized on two mobile genomic islands.* Int J Med Microbiol, 2008. **298**(5-6): p. 411-28.
44. Tseng, T.T., B.M. Tyler, and J.C. Setubal, *Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology.* BMC Microbiol, 2009. **9 Suppl 1**: p. S2.
45. Okuda, S. and H. Tokuda, *Lipoprotein sorting in bacteria.* Annu Rev Microbiol, 2011. **65**: p. 239-59.
46. Beckwith, J., *The Sec-dependent pathway.* Res Microbiol, 2013. **164**(6): p. 497-504.
47. Chatzi, K.E., et al., *Breaking on through to the other side: protein export through the bacterial Sec system.* Biochem J, 2013. **449**(1): p. 25-37.
48. Palmer, T. and B.C. Berks, *The twin-arginine translocation (Tat) protein export pathway.* Nat Rev Microbiol, 2012. **10**(7): p. 483-96.
49. Stanley, N.R., T. Palmer, and B.C. Berks, *The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in Escherichia coli.* J Biol Chem, 2000. **275**(16): p. 11591-6.
50. Robinson, C., et al., *Transport and proofreading of proteins by the twin-arginine translocation (Tat) system in bacteria.* Biochim Biophys Acta, 2011. **1808**(3): p. 876-84.
51. Cain, J.A., N. Solis, and S.J. Cordwell, *Beyond gene expression: The impact of protein post-translational modifications in bacteria.* J Proteomics, 2014. **97**: p. 265-86.
52. Waridel, P., et al., *Evidence for a new post-translational modification in Staphylococcus aureus: hydroxymethylation of asparagine and glutamine.* J Proteomics, 2012. **75**(6): p. 1742-51.
53. Tanaka, K., S.I. Matsuyama, and H. Tokuda, *Deletion of lolB, encoding an outer membrane lipoprotein, is lethal for Escherichia coli and causes accumulation of lipoprotein localization intermediates in the periplasm.* J Bacteriol, 2001. **183**(22): p. 6538-42.
54. Seydel, A., P. Gounon, and A.P. Pugsley, *Testing the '+2 rule' for lipoprotein sorting in the Escherichia coli cell envelope with a new genetic selection.* Mol Microbiol, 1999. **34**(4): p. 810-21.

55. Hara, T., S. Matsuyama, and H. Tokuda, *Mechanism underlying the inner membrane retention of Escherichia coli lipoproteins caused by Lol avoidance signals.* J Biol Chem, 2003. **278**(41): p. 40408-14.

Chapter 2. Genetic analysis of the *Legionella pneumophila* strain Corby Diversity Generating Retroelement.

Abstract

Diversity-generating retroelements (DGRs) are a unique family of retroelements that confer selective advantages to their hosts by facilitating localized DNA sequence evolution using a specialized error-prone reverse transcription process. We characterized a chromosomal DGR in *Legionella pneumophila* (*Lp*), an opportunistic human pathogen and etiological agent of Legionnaires' disease. The *Lp* DGR is found within a horizontally acquired genomic island and it can theoretically generate 10^{26} unique nucleotide sequences in its target protein encoding gene, *ldtA*, which upon translation would create a repertoire of 10^{19} distinct peptide sequences within the 3' C-type lectin domain (Clec) of LdtA. Expression of the *Lp* DGR resulted in transfer of DNA sequence information from a template repeat (TR) to a variable repeat (VR) accompanied by adenine-specific mutagenesis of progeny VRs found at the 3' end of *ldtA*. We have analyzed the DNA, RNA, and protein requirements for mutagenic homing in *Lp* and found they are analogous to those of the *Bordetella bronchiseptica* (*Bb*) phage. *Lp* contains a hierarchical regulatory circuit controlled by a few master regulators which coordinate progression through its biphasic lifecycle. Activity of the Corby DGR appears to be regulated by these global networks as mutants have increased levels of mutagenic homing. This work suggests that mutagenic homing in the *Bb*, *Lp*, and perhaps all DGRs functions through a conserved mechanism, likely template dependent reverse transcription (TPRT).

Introduction

Diversity-generating retroelements benefit their hosts by accelerating the evolution of target proteins (TP) ligand binding domains [1-3]. DGRs were discovered in a *Bordetella bronchiseptica* bacteriophage, BPP, which uses site-specific, error-prone reverse transcription to generate diversity in a gene that encodes tail fibers responsible for host ligand recognition [1]. The process of gene diversification requires a DGR-encoded reverse transcriptase (RT), an accessory variability determinant (Avd or an equivalent protein), and a template repeat (TR)-derived RNA intermediate [1, 4]. The TR-RNA provides a template for reverse transcription, during which TR adenine residues are copied into any of the four nucleotides. This diversified cDNA displaces a variable repeat (VR) located at the 3' end of the TP encoding gene [4]. This unidirectional, targeted, mutagenic retrotransposition process is called mutagenic homing. Target recognition of TP encoding genes requires two *cis*-acting sequences at the 3' end of VR; the initiation of mutagenic homing (IMH) element and a DNA hairpin/cruciform structure [5]. Mutagenic homing operates through a copy, diversify, and replace mechanism during which *cis*- and *trans*-acting factors required for further rounds of diversification are preserved, allowing iterative optimization of TP function. Structural studies of BPP demonstrated that TR adenines are precisely positioned to correspond to residues in the ligand binding pocket of a C-type lectin (CLec) domain at the C-terminus of the tail fiber protein, reflecting co-evolution between the genetic mechanism that generates diversity and the protein scaffold that displays it [6].

The *Bordetella* BPP phage DGR has provided the sole paradigm for mechanistic studies for this family of retroelements. Bioinformatic analysis of deposited nucleotide

database, metagenomic datasets, and whole-genome shotgun contigs revealed that DGRs are widespread throughout the bacteria domain, with representatives in every phylum with significant sequence coverage [7, 8]. DGRs are enriched within a few phyla, *Firmicutes, Proteobacteria*, as well as *Bacteroidetes*, however it is unclear if this enrichment has biological significance or is due to sequencing bias. Interestingly, DGRs have also been identified within an archaeal species, *Nanoarchaeum equitans*, extending the distribution of these elements outside of the bacterial domain (Personal communication Blair Paul).

Bioinformatic analysis identified DGRs within several strains of the opportunistic human pathogen *Legionella pneumophila* (*Lp*), a gram-negative intracellular pathogen which has evolved strategies to evade predation in the environment, leading to accidental virulence in humans [9, 10]. These elements encode nearly identical diversification machinery yet each has a VR with a unique pattern of adenine mutagenesis suggesting individualized diversification in response to environmental pressures. We determined the necessity of *Lp* DGR conserved genes and factors to mutagenic homing and found they were similar to the *Bordetella* phage, suggesting DGRs may function through a conserved mechanism. Observations of increased mutagenic homing during *Lp* transition from replicative to transmissive state suggested DGRs might be controlled by host global regulatory networks. This hypothesis was evaluated by generating deletions in key regulators, *relA* and *spoT* [11]. Analysis of these mutants revealed that certain DGR components are regulated and that the abundance of TR-RNA transcripts controls levels of mutagenic homing. To verify the increased levels of mutagenesis in regulatory mutants was not due to homologous

recombination into VR we determined the contribution of host RecA-dependent recombination machinery to mutagenic homing and found the levels unaffected. Cumulatively, this suggests that *Lp* DGRs have integrated into host global regulatory systems associated with stress. Finally, we found that all *Lp* DGR containing strains are capable of supporting mutagenic homing, supporting our hypothesis that these elements are being maintained and likely selected for. This work is the first functional analysis of a bacterial chromosomal DGR, using *Lp* as a model system.

Results

Anatomy of a *Legionella* DGR. The DGR in Figure 1A is located within a ten kilobase pair (kbp) genomic island within the chromosome of *Lp* strain Corby. Genes flanking the genomic island are predicted to encode a heavy metal transport system (*helA-C*) and a type IV secretion system with associated regulators (*lvrA-C*). The island shows signs of recent horizontal acquisition, as indicated by G+C content (45%) that differs from the rest of the genome (38%), the presence of a transposase and an unrelated RT [12]. Interestingly, the 10 kbp DGR-containing island is incorporated into a larger (64 kbp) Integrative and Conjugative Element (ICE) [13]. The retroelement itself encodes a DGR-type RT [1], an Avd homolog, cognate TR and VR sequences that differ at sites corresponding to adenines in TR, and tandem stem-loop/cruciform structures downstream of VR [4, 5]. The 148 bp TR contains 43 adenines which most often occupy the first two positions of AAC or AAT codons, allowing maximal amino acid diversity while excluding the possibility of nonsense mutations generated by adenine-

mutagenesis (Figure. 1B). Following mutagenic homing, the *Lp* TR can theoretically generate 4^{43} (~10^{26}) unique DNA sequences capable of encoding ~10^{19} different polypeptides, a repertoire of massive proportions. VR encoded sequences are located at the C-terminus of LdtA, within a domain predicted to adopt a CLec fold similar to the distal end of the BPP phage tail fiber protein, Mtd (Sup. Figure 1) [14]. LdtA is predicted to contain a bipartite N-terminal signal peptide where the first domain is a twin-arginine translocation (TAT) motif that differs from the consensus in *Escherichia coli* (*E. coli*) but is characteristic of known and putative TAT substrates in *Lp* [15, 16]. The second domain is a lipobox motif that is also non-canonical from those found in *E. coli* [17]. The TAT pathway is an alternative secretion system found in plants and bacteria that can translocate folded proteins or complexes across membranes [16] and lipobox motifs mediate signal peptide cleavage, lipid modification, and anchoring to the inner or outer membrane [18]. Although the ability of TAT and lipobox secretion motifs to function in concert has not been entirely characterized [19, 20], we hypothesized that the N-terminus of LdtA mediates secretion, membrane localization, and potentially surface exposure.

The *Lp* DGR is a functional retroelement. As a first step in characterizing the *Lp* DGR, we determined if it is capable of mutagenic homing. A polymerase chain reaction (PCR) based assay was used to detect RT-dependent transfer of an invariant sequence tag from TR to VR, and adenine mutagenesis was evaluated by sequencing amplified retrohoming products (Figure. 2). For this assay *avd, RT*, and a modified TR containing a 20 bp tag consisting of G+C residues (TR-GC) was expressed *in trans* to the wild type DGR on a plasmid vector (pATR, Figure 2B). Negative controls included

constructs expressing wild type TR (no tag) or a catalytically inactive RT (pSMAA). Following induction, total DNA was extracted and homing products were amplified using primer sets that annealed to the GC tag and sequences upstream (P1/P3) or downstream (P2/P4) of VR. Homing products were readily detected, identified by transfer of the tag from TR-GC to VR (Figure. 2C), and sequencing of GC-tagged VRs revealed adenine-specific mutagenesis (Sup. Figure 2A). These results show that the *Lp* Corby DGR encodes functional components that are capable of catalyzing adenine-specific mutagenic homing to the VR of *ldtA*.

In the experiment in Figure 2C, efficient transfer of the GC tag was detected following expression of *avd*, TR-GC, and *RT in trans* to the chromosomal DGR. To measure activity of the native element under endogenous conditions, we inserted the 20 bp GC tag into the chromosomal TR by allelic exchange and assayed mutagenic homing by PCR. Transfer of the GC tag and adenine mutagenesis were both observed (Sup. Figure 2B, C) but detection required an increase in the number of PCR amplification cycles compared to the experiment in Fig. 2B (25 vs. 35 cycles). The low level of activity of the native chromosomal element was not surprising. DGR-mediated mutagenesis is stochastic and the vast majority of diversified target genes are likely to encode inactive products. Mutagenic homing is expected to be tightly controlled to prevent a loss of fitness due to over-diversification (see *Discussion*).

***cis*- and *trans*-acting factors required for mutagenic homing.** Our understanding of mutagenic homing derives almost exclusively from studies using the *Bb* BPP DGR [1, 5]. We were curious to determine if observations with the phage are

applicable to other DGRs, especially those encoded on bacterial chromosomes. To explore requirements for mutagenic homing by the *Lp* DGR, *avd*, TR, and *RT* were deleted *en bloc* from the *Lp* genome and pATR (Figure 2B), or derivatives with mutations in *trans*-acting factors, were tested in our PCR-based assay. The ability to detect homing products required the TR tag and expression of an active RT, and homing was also dependent on expression of *avd* (Figure 2D). *avd* is predicted to encode a small, basic protein (14.1 kDa, pI 9.3) with homologs in other DGRs, including BPP, where it serves an essential function and has been shown to form a positively charged pentameric barrel that interacts with RT [1, 21, 22]. IMH*, located at the 3' terminus of TR, is a 32 bp sequence with 2 mismatches to IMH. In BPP, IMH and IMH* are essential components that are predicted to facilitate assembly of a complex between VR, TR-RNA, and RT during priming and reverse transcription [4, 5]. Deletion of IMH* on the pATR donor plasmid or IMH on the recipient chromosome abrogated activity of the *Lp* DGR (Figure 2D).

In BPP, target site recognition is also dependent on the presence of an inverted repeat downstream of IMH that forms a hairpin/cruciform structure in supercoiled DNA [5]. These elements are highly conserved in phage DGRs where they consist of 7-10 bp GC-rich stems and 4 nt loops. A divergent yet potentially analogous element, composed of two tandem repeats with GC-rich stems of different lengths and identical 3 nt loops, are present downstream of the *Lp ldtA* IMH (Figure. 2B). To examine their role in homing, we generated mutants in which sequences comprising the 3’ halves of either stem were replaced with complementary nucleotides to disrupt base pairing. As shown in Figure 2E, disruption of stem-loop 1 (SL1) or stem-loop 2 (SL2) eliminated homing.

Our results show that the *Lp* and BPP DGRs operate in a fundamentally similar manner, using conserved *cis*- and *trans*-acting components for adenine-specific mutagenic homing.

Host global regulatory systems affect levels of mutagenic homing. Bacteria live in dynamic environments and constantly monitor their environment to regulate physiological processes such as metabolic pathways or regulation of virulence mechanisms. Observations of starved *E. coli* identified the cessation of rRNA synthesis with the production of an alarmone, guanosine 5'-diphosphate-3'-diphosphate (ppGpp), and this response to harsh environments coupled with alarmone synthesis has been called the stringent response [11]. ppGpp is synthesized from guanosine diphosphate (GDP) or guanosine triphosphate (GTP) as well as adenine triphosphate (ATP) and can be hydrolyzed to GDP/GTP and a pyrophosphate. In *E. coli*, cellular levels of ppGpp are typically controlled by two enzymes; *relA* has synthases activity while *spoT* has synthase and hydrolase activity [11]. *Lp* progresses through phenotypically distinct states by controlling intracellular levels of ppGpp through genes homologous to *relA* and *spoT* [11, 23]. *Lp* senses deteriorating conditions within its host and in response synthesizes ppGpp and these increased levels of alarmone triggers signal transduction cascades that regulate over 50% of genes, typically resulting in the transition from the non-flagellated, replicative state to a motile, transmissive state [23, 24].

Low levels of mutagenic homing were observed in *wt* Corby cells grown to stationary phase (Sup. Figure 2B) and this is typically when levels of ppGpp coordinate *Lp* transition from replicative to transmissive phase. We hypothesized mutagenic

homing might be regulated by activity of RelA and SpoT. To detect DGR activity a 20bp "GC" invariant tag was introduced into the chromosomal TR of *wt* Corby (TRwTAG) cells (Figure 3A). In-frame, unmarked deletions of *relA* (TRwTAGΔ*relA*) and *spoT* (TRwTAGΔ*relA*Δ*spoT*) were generated using allelic exchange with *sacB* counter-selection, a technique often used in *Lp* [25]. Cells were grown in rich media which accurately replicates the biphasic lifecycle of *Lp* and samples were taken at optical densities (OD) corresponding to changes in phenotypic state [23]. Total DNA was extracted and subject to PCR using primers annealing to the invariant tag and to flanking nucleotide sequences in either *ldtA* or *avd* (Figure 3B). Controls included untagged cells (*wt*) and double mutant cells with deletion of the DGR RT TRwTAGΔ*RT*Δ*relA*Δ*spoT*. Transfer of the invariant tag could be detected in TRwTAGΔ*relA* or TRwTAGΔ*RT*Δ*relA*Δ*spoT* cells grown to early stationary phase (Sup. Figure 3) and these levels increased as cells reached late stationary phase (Figure 3C). In agreement with earlier work, very low levels of mutagenic homing could be detected in TRwTAG cells grown to late stationary phase. Transfer of the tag was not detected in *wt* or TRwTAGΔ*RT*Δ*relA*Δ*spoT* cells (Figure 3C and Sup Figure 4). Sequencing of PCR products from mutant cells from all time points revealed transfer of the invariant tag and adenine mutagenesis verifying mutagenic homing (Sup. Figure 4). This suggests that activity of RelA and SpoT which are responsible for the modulation of global regulatory programs also, at least partially, regulate levels of mutagenic homing.

Analysis of *Lp* DGR genes during its biphasic lifecycle. In *Lp* the regulatory cascade controlled by *relA* and *spoT* is hierarchical, involving two component regulatory systems, RNA binding proteins as well as many other factors and modulation of gene

expression occurs transcriptionally and post-translationally [11, 23, 26]. We wanted to assess if differences in levels of mutagenic homing observed in TRwTAGΔ*relA* and TRwTAGΔ*relA*Δ*spoT* cells as compared to TRwTAG cells resulted from alterations in expression of DGR genes or factors. Wild type and mutant *Lp* cells (Δ*relA*Δ*spoT*) were grown to ODs corresponding to early exponential, early or late stationary phase (as above), total RNA extracted, cDNA libraries synthesized and changes in RNA transcript number were analyzed using reverse transcriptase-PCR (RT-PCR) with primers specific to DGR components (*TP*, *avd*, *RT*, or TR) or control genes (*recA*, *fliC*, and *rpoS*). Controls included Corby cells with deletions in the RelA/SpoT regulated sigma factor responsible for activation of the flagellar regulon (Δ*fliA*) and DGR adjacent genes with homology to T4SS regulators (Δ*lvrRABC*) [11, 13, 27]. PCR analysis of *wt*, *ΔfliA*, and *ΔlvrRABC* cells extracted from early exponential (EE), early stationary (ES), and late stationary (LS) phase revealed changes in expression of control genes consistent with *Lp* progression through its biphasic lifecycle (Figure 4) [24]. While expression of the stationary phase sigma factor *rpoS* remains stable regardless of growth phase, expression of *fliC* (encodes for flagellin) increased in *wt* and Δ*lvrRABC* cells as they progress from the replicative to transmissive state but this increase in expression is not observed in Δ*fliA* or Δ*relA*/Δ*spoT* cells which is consistent with *filA* as a sigma factor and the *relA*/*spoT* regulatory cascade. Interestingly, while expression of TR-RNA increases in Δ*relA*/*ΔspoT* cells during late stationary phase there does not appear to be any change in *RT* or *avd* expression (data not shown). *TP* expression in Δ*relA*/Δ*spoT* cells appears to be higher than in *wt*, Δ*fliA*, or Δ*lvrRABC* cells. This suggests that

certain DGR components are regulated by levels of ppGpp although the precise mechanism is unclear.

The contribution of RecA-dependent DNA recombination machinery to mutagenic homing. While previous studies in BPP revealed that mutagenic homing was a *recA* independent process [4], we wanted to assess the possibility that increased levels of homing in Corby*ΔrelAΔspoT* cells represented diversified cDNA products being recombined into the chromosome by host machinery independently of TPRT. A deletion of *recA*, which encodes a necessary bacterial DNA recombination protein, was introduced into *Lp* TRwTAG*ΔrelAΔspoT* cells by allelic exchange, as above. The contribution of RecA to DGR mutagenic homing was assessed in cells grown to ODs corresponding to EE, ES, or LS growth phase and transfer of the invariant tag into VR was detected by PCR (as above). *Lp* TRwTAG*ΔrelAΔspoT* and TRwTAG*ΔrelAΔspoTΔrecA* cells demonstrated similar levels of mutagenic homing suggesting RecA has no effect on mutagenic homing (Figure 5).

Overexpression of TR-RNA effects levels of mutagenic homing. We wanted to assess if the increase in TR-RNA transcript numbers was solely responsible for the alteration in mutagenic homing observed in TRwTAG*ΔrelAΔspoT* cells. The DNA regions surrounding the Corby TR, including portions of *avd* and *RT*, were analyzed and a rho-independent transcriptional terminator was identified approximately 500 bps from the start codon of *RT* (Figure 6A) [28]. A DNA fragment from the stop codon of *avd* to this predicted transcriptional terminator was cloned into the inducible plasmid pMMB208 (generating pTR) and transformed into *Lp* TRwTAG cells. Cells were grown in rich

media and induced for expression of pTR for 0, 2, or 4 hours and total DNA was analyzed by PCR using primers specific to the invariant tag and flanking DNA sequences, as above. In cells induced for 2 (Figure 6B) and 4 (data not shown) hours transfer of the invariant tag could be detected in pTR expressing cells but not for empty vector expressing cells however, the level of mutagenic homing appears to be far less than cells expressing pATR (data not shown) as well as *relA*/*spoT* mutant cells.

Mutagenic homing in related *Lp* strains. An interesting observation regarding the distribution of DGRs within the bacterial domain is that representative elements are found in many phyla but a DGR containing phylum may not have elements in every species [7]. To address the distribution of DGRs within the *Legionella* genus we assayed for and identified several putative elements which are highly homologous to the Corby element within *Lp* strains D5549, D5572, and D5591 as well as *L. tunisiensis* strain LegM, discussed in greater detail in chapter 3. The *Legionella* DGRs encode nearly identical diversification machinery: stem loop/cruciform structures, *avd*, *TR*, as well as *RT* and we took advantage of this to determine if these related strains/species were capable of supporting mutagenic homing. Plasmids pATR and pSMAA were transformed into D5572, D5591, and Corby. Controls included *wt* Corby cells and plasmids not transformed into *Lp*. Using target gene sequence specific primers (N1F for Corby and D5572 or N2F for D5592) as well as a reverse primer specific for the invariant tag (N1R), transfer of the invariant tag was detected for all three strains expressing pATR but not for cells expressing pSMAA, *wt* Corby cells, or plasmid only controls (Figure 7). Sequencing of PCR products revealed transfer of the invariant tag and adenine mutagenesis, confirming mutagenic homing (data not shown).

We then assessed if all proteobacteria are capable of supporting mutagenic homing by expressing DGRs *in trans* within two organisms without identifiable elements. The first was *Lp* strain Philadelphia containing a nearly identical ICE as Corby except that it is lacking a DGR [13] and the second was the wild type *E. coli* strain MG1655 [29]. Both strains were transformed with pATR or pSMAA and analyzed for transfer of the invariant tag from plasmid TR to a VR found on a second plasmid (pVR). This two plasmid system had been shown capable of supporting mutagenic homing in Corby (data not shown). While transfer of the invariant tag could be detected by PCR in Philadelphia, it could not be detected in MG1655 or in additional *E. coli* strains tested (data not shown). This suggests *Lp* strains may contain additional factors necessary for mutagenic homing or necessary factor(s) in *Lp* and *E. coli* has diverged sufficiently to allow homing in one but preclude it in the other.

Discussion

To date, over 300 unique DGRs have been identified in phage, plasmid, or bacterial genomes and are associated with an array of diverse ecological niches. Despite their widespread distribution in nature and capacity to confer selective advantages, only a single, phage-encoded DGR has been studied in mechanistic detail [1, 5, 6, 21]. Our discovery of a functional DGR in *Lp* provides a bacterial system for comparative analysis. We demonstrate that targeted, adenine-specific mutagenesis occurs in *Lp* providing direct support for the hypothesis that this is a capability common to all DGRs. Both the *cis*- and *trans*-acting requirements for DGR activity in *Lp* are

analogous to those in *Bordetella* phage, highlighting the conserved nature of the mutagenic homing mechanism.

To our knowledge, the amount of diversity that can be generated by *Lp* DGRs is greater than for any characterized biological system, including the diversification of immunoglobulin scaffolds during mammalian immune responses [30]. The comparison between DGR-mediated diversity and the generation of immunity is instructive in several ways. In both cases, the genetic mechanisms responsible for creating diversity have co-evolved with protein scaffolds to display it, the immunoglobulin fold for antibodies and T-cell receptors and the CLec fold for DGRs. Additionally, the same basic sequence of target-gene diversification, surface display of variable proteins, and selection leading to amplification appears to hold in either case. DGRs, however, operate under a unique constraint. Although mutagenesis is highly directed, it is inherently stochastic and the vast majority of mutagenic events are likely to be deleterious. For TPs that have function, we predict the frequency of mutagenic homing will be low, or regulated, to balance the loss of fitness resulting from mutagenesis with the advantages conferred by accelerated evolution. This may explain the low level of basal activity observed for the native *Lp* DGR and our ability to increase mutagenesis by exogenous expression of *trans*-acting components.

Several lines of evidence support a hypothesis that *Lp* DGRs, or at least certain components, are regulated by host systems. While a low level of basal activity is observed in *wt* Corby cells, double deletion mutants of *relA* and *spoT* demonstrate increased activity that is most detectable in cells entering stationary phase. This

typically represents the transition from replicative to transmissive state [11] that occurs through the sensing of cellular levels of the alarmone ppGpp. ppGpp can directly modulate gene expression through the DNA binding protein DksA, through altering RNA polymerase (RNAP) affinity for gene promoters as well as through indirectly modulate gene expression via sigma factor competition of the housekeeping sigma factor, σ^{70} [11]. The observation that *wt*, *ΔrelA*, and Δ*relA*Δ*spoT* display low, medium, and high levels of mutagenic homing respectively further supports the hypothesis that intracellular levels of ppGpp are influencing DGR activity. The increase in mutagenic homing in *relA*/*spoT* mutants appears to occur through TPRT and does not involve an increased synthesis of diversified cDNA followed by homologous recombination, as demonstrated by the *recA* mutants. Additionally, analysis of the *relA*/*spoT* double deletion mutants identified an increase in TR-RNA transcript number. However, this increase in transcript number was only observed for TR as levels of *avd* as well as *RT* were constant, regardless of growth phase. This is of particular interest when considering that several gene products and factors are necessary for mutagenic homing. The most parsimonious explanation for these observations is that altered rates of mutagenic homing result from changes in TR-RNA transcript number and this is partially supported by overexpression of TR-RNA in TRwTAG cells. These observations are consistent with TR being a direct target of regulation by ppGpp however we cannot rule out the possibility of post-transcriptional or post-translational regulation [23]. Additionally, the *relA*/*spoT* regulatory cascade intersects with other independent regulatory circuits and multiple regulatory cascades might be functioning to either activate or repress DGR activity. The observation that increased levels of mutagenic

homing are only detected during certain growth phases is consistent with DGRs being regulated by multiple pathways. One regulatory network that functions independently of *relA* and *spoT* is the regulation of many *dot*/*icm* T4SS effector genes during late transmissive phase by the two component regulatory system *pmrA*/*B* [31]. This system is thought to prime the bacterium for invasion of host cells in an anticipatory fashion and is active during a similar time as the observed increase in mutagenic homing. While the exact mechanism of regulation still needs to be determined, we propose that the Corby DGR has integrated the regulation of the diversification machinery into global systems which respond to stress in order to maximize the benefit of target gene diversification.

We have identified several homologous DGRs in at least two species of *Legionella*. These elements are found within genomic islands that are generally incorporated into larger ICE that have been demonstrated to be horizontally transferred between *Lp* strains and *Legionella* species which would provide a means for distribution throughout bacterial populations [13]. We demonstrated that every *Lp* strains tested was capable of supporting mutagenic homing. However, not every bacterium is capable of supporting mutagenic homing as ectopic expression of DGR components in several strains of *E. coli* resulted in no detectable transfer of sequence information. Furthermore, expression of TR-RNA is an integral step in initiating mutagenic homing and expression of DGR components in *E. coli* does not generate TR-RNA (personal communication-Huatao Guo). While *Lp* and *E. coli* share the same phylogenetic order they are found in different classes, raising the possibility that missing or divergent host factors may play a role in discriminating organisms capable of supporting DGR mutagenic homing. One could imagine a situation where a DGR containing mobile

element invades a cell which lacks the requisite host machinery for mutagenic homing thus negating any DGR-derived benefit.

DGRs are found widely distributed within the bacterial domain and have been demonstrated to facilitate the accelerated evolution of DNA nucleotide sequences of target protein ligand binding domains [3]. This work represents the first characterization of a bacterial, chromosomally encoded DGR. We have verified the Corby element is active and this work suggests all DGRs function through a conserved mechanism currently proposed to be TRPT. The distribution of homologous DGRs throughout *Legionella* suggests these elements are being maintained and that diversification of TPs is regulated by host global systems and likely coincides with environmental stresses.

Materials and Methods

Bacterial Strains, Growth, and Mutant Construction. *Lp* Corby and all other referenced strains [32] were a kind gift from Dr. Natalia Kozak (CDC). *Lp* Corby and derivatives were routinely maintained in culture in yeast extract (PYG) broth or on buffered charcoal-yeast extract (BCYE) media as previously described [33]. In-frame deletions and substitution mutations were constructed using allelic exchange with the *sacB* negative selection marker on BCYE agar containing 7.5% sucrose [34]. *Lp* Corby gene loci targeted for mutational analysis included *tatB* (LpC_3208), *relA* (LpC_0872), *spoT* (LpC_1492), *recA* (LpC_1245), *ldtA* (LpC_1853), *avd* (LpC_1854), *RT* (LpC_1855) and intergenic regions between *ldtA* and *avd* or *avd* and *RT* representing stem/loops, VR, or TR, respectively. The broad host vector pMMB208 was used for

complementation of mutants and protein overexpression and cells transformed with this vector were grown in media supplemented with 5µg/mL chloramphenicol (Cm) [35].

Plasmid Construction and Mutagenesis. pATR (and derivative donor plasmids) for DGR homing assays was constructed by cloning sequences extending from 75 bps upstream of LpC_1854 to the stop codon of LpC_1855 into the broad host range vector pMMB208 and inserting a 20 bp GC tag into position 98 of TR; all donor plasmids are derivatives of this initial pATR construct. The catalytically inactive RT derivative (Figure 2) contains a mutation replacing the essential RT amino acid motif YVDD with SMAA [4]. The IMH* deletion derivative is missing sequences from TR position 108-140. The *avd* mutant carries a deletion that removes all sequences between the *avd* start and stop codons. For studies involving overexpression of the Corby TR-RNA, a DNA fragment from the stop codon of *avd* to 570 bps downstream from the start of *RT* was cloned into the plasmid pMMB208 to generate pTR. *Lp* cells were transformed with pTR and selected for on BCYE + Cm plates. Cells were grown to the indicated OD and expression of TR-RNA was induced by the addition of IPTG, to a final concentration of 1mM, to the growth media.

PCR-based DGR Homing Assays. Homing assays were performed as previously described with minor modifications [4, 5]. In short, *Lp* Corby cells harboring pATR, or mutant derivatives, were sub-cultured in PYG broth supplemented with Cm to an OD_{590} of 0.2. Cells were grown for four hours and induced for DGR component expression by the addition of IPTG to a final concentration of 1mM for four hours. Cells were harvested and DNA extracted by a commercial kit (Qiagen). Polymerase chain

reaction (PCR) was used to detect transfer of the invariant tag from donor plasmid TR or chromosomal TR sequences to chromosomal VR sequences using the following primer pairs: P1- GCGGCATTGACGGATGAGCC, P2- ACAGGAACACAAACGCAGAC, P3- GTCTGCGTTTGTGTTCCTGT, P4- GCTTCATCTCGACACACAGGCGAATTTCC, and P5- CATGATTCTGGCTTTCGGCTGGCATTACG. Amplified products were cloned (Invitrogen-TOPO) and sequenced to verify transfer of the tag from TR to VR and to detect adenine mutagenesis.

Reverse transcription-PCR. Various *Lp* strains were grown in PYG to the indicated ODs where samples were taken and frozen in a dry-ice ethanol bath in order to stop degradation of RNA. Total RNA was extracted from samples using a commercially available kit (Ambion) and used as a template for cDNA synthesis (Invitrogen). DNA concentration was determined by spectrophotometer and normalized by concentration. PCR using gene specific primers was performed and products visualized.

Figure Legends

Figure 1. A *Legionella pneumophila* DGR. (*A*) A DGR in *Lp* strain Corby is found within a ten kbp genomic island (green dashed box) adjacent to a heavy metal transport gene cluster (*helA-C*) and Type IV secretion system regulatory genes (*lvrA-C*). The DGR-containing genomic island has a higher G+C content (percentile score) than flanking regions, a non-DGR RT (dark blue), genes annotated as hypothetical proteins (grey), and putative transposase (orange). DGR loci are expanded with the TP gene (*ldtA*), with a predicted bipartite secretion peptide (green box with colored boxes) with TAT and lipoprotein motifs (Lpp), accessory protein gene (*avd*), and DGR-encoded reverse transcriptase (*RT*) identified. Additional DGR elements: DNA stemloop/cruciform structure (underline pink), IMH, and IMH* are indicated. Blue arrows represent mutagenic homing as transfer of nucleotide sequence information from TR to VR via a TR-RNA intermediate and the adenine mutagenesis is depicted as the change of TR-RNA adenines into any nucleotide (N) in cDNA. The LdtA signal peptide contains a polar N-region (residues 1-11), followed by a hydrophobic core (residues 12-15), and a non-polar C-region (residues 17-23, not shown). (*B*) Alignment of *Corby* DGR VR/TR shown in the *ldtA* reading frame. TR adenines (red) usually occupy the first two positions of AAC or AAT codons (shaded) and correspond to substitutions in VR. Nucleotide sequences downstream of the VR stop codon representing DNA stemloop/cruciform structure (underline pink) structures are shown.

Figure 2. The *Lp* DGR is capable of mutagenic homing. (*A*) *Lp* Corby DGR cassette with components indicated as in Fig. 1A. (*B*) Experimental design. pATR encodes *avd*, TR with 20 bp GC PCR tag (pink cylinder), and *RT* expressed from P*tac*. DGR-mediated transfer incorporates the GC tag into the chromosomal VR. Primer binding sites (P1-P5) for PCR-based mutagenic homing assays are shown and deletions removing IMH or disruption of the first (MSt1) or second (MSt2) stem/loop structure are indicated by blue brackets. (*C*) DGR homing assays. Donor plasmids (pATR and derivatives) were transformed into wild type *Lp* Corby, P*tac* expression induced, and genomic DNA extracted and used for PCR with primers shown in *B*. Donor pATR derivatives encoded wild type (*wt*) or mutant (*RT*-) RT alleles, with (+) or without (-) TR tags. Equivalent amounts of template DNA were used for PCR assays (P1+P4). Non-specific PCR amplification of donor plasmid indicated by red asterisk. (*D*) pATR constructs used in (*C*) and additional derivatives with deletions removing *avd* (*avd*-) or IMH (IMH-) were transformed into *Lp* Corby deleted for sequences from *avd* to *RT* (*Δavd-RT*) with IMH present (*wt*) or deleted (*IMH*-). DGR homing assays were conducted as in (*C*) with equal amounts of template DNA (P1+P5). Non-specific PCR amplification of donor plasmid indicated by red asterisk (*E*). pATR donor plasmids with wild type (*wt*) or mutant RT alleles (*RT*-) were transformed into *Lp* Corby *Δavd-RT* with wild type (*wt*) or disrupted stem/loop structures (MSt1, MSt2). DGR homing assays were conducted as in (*C*) with equal amounts of template DNA (P1+P5).

Figure 3. Altered rates of mutagenic homing in *relA*/*spoT* mutant *Lp*. (*A*) Experimental design showing insertion of the 20 bp invariant, GC tag (as in Figure 2) into the chromosomal TR. Blue arrows represent mutagenic homing as transfer of nucleotide sequence information from TR to VR via a TR-RNA intermediate. The adenine mutagenesis is depicted as the change of TR-RNA adenines (red A) into any nucleotide (black N) in cDNA. (*B*) Expanded view of 3' *ldtA*, IMH, stemloops/cruciform structure, and *avd*. Specific primers and relative binding sites are indicated. (*C*) Various *Lp* TRwTAG strains were grown to late stationary phase, total DNA extracted, and assayed for transfer of invariant TAG from TR to VR using indicated primers. Presence (+) or absence (-) of DGR *RT*, *relA*, and *spoT* is indicated. Primers TAGFor/avdRev detect transfer of TAG to VR as a surrogate for mutagenic homing, with ldtAFor/avdRev showing equal loading of DNA.

Figure 4. Expression of DGRs genes in various *Lp* mutants. Various mutants of *Lp* were grown to ODs which correspond to growth phases: early exponential (EE), early stationary (ES), or late stationary (LS). Total RNA was extracted and cDNA or mock libraries constructed (see Materials and Methods) and changes in transcript numbers assayed by PCR to gene specific primers: *recA* (DNA recombination), *fliC* (flagellin), *rpoS* (stationary phase sigma factor), *TP* (DGR TP ldtA), *RT* (DGR RT), and TR (DGR template repeat). In-frame deletions of various regulatory genes were made (see Materials and Methods): the global regulators *relA* and *spoT*, the flagellar sigma factor *fliA*, or local T4SS like regulators *lvrR*-*C*.

Figure 5. Deletion of RecA-dependent host machinery has no effect on mutagenic homing. An in-frame deletion of *recA* (Δ*recA*) was introduced into *Lp* TRwTAGΔ*relA*Δ*spoT* (ΔΔ) cells. Double and triple mutant strains were grown up to ODs corresponding to growth phases: early exponential (EE), early stationary (ES), or late stationary (LS). Total DNA was extracted and assayed for transfer of the invariant TAG from TR to VR using primers TAGFor/avdRev while primers ldtAFor/avdRev ensured equal loading of DNA. Similar levels of detection with TAGFor/avdRev indicate deletion of *recA* has no effect on mutagenic homing.

Figure 6. Overexpression of TR-RNA alters levels of mutagenic homing. (*A*) Analysis of the Corby DGR for transcriptional terminators using ARNoLD program [28] identified two putative Rho-independent transcriptional terminators. These terminators are on the minus strand, the first is found within the two stem loop/cruciform structures and the second within the 5' of *RT*. Stems are indicated in blue, loops in red, and free energy of stem-loop region indicated in kcal/mol. (*B*) *Lp* Corby cells expressing either empty vector (EV) or pTR were induced for two hours, total DNA extracted, treated with RNase, and assayed with primers to detect transfer of TAG from plasmid TR to chromosomal VR (TAGFor/avdRev) or to ensure equal loading of DNA (ldtAFor/avdRev). For comparison, Corby TRwTAGΔ*relA*Δ*spoT* +/- DGR RT and Corby TRwTAGΔ*relA* +/- DGR RT grown to stationary phase were assayed using similar conditions.

Figure 7. DGR containing clinical strains of *Lp* are capable of supporting mutagenic homing. Analysis of a library of *Lp* identified a number of elements with putative DGRs which are highly homologous to the element in Corby. Two *Lp* clinical isolates, D5572 and D5591, were transformed with the plasmid vector over-expressing Corby *avd,* TR-GC, and *RT* (pATR), as in Figure 2, *in trans* to their native elements. Cells were induced for plasmid expression with the addition of IPTG, total DNA was extracted, and assayed for transfer of invariant TAG from plasmid TR to chromosomal VR using primers N1F/N2F (similar to ldtAFor) and N1R (similar to TAGRev). Controls include pATR where the RT has been catalytically inactivated (pSMAA) and PCR of plasmids extracted from *Lp* without induction.

Sup. Figure 1. Predicted structures of the C-terminal domains of LdtA, LdtB, and LdtC in ribbon representation. α-helices (red), β-strands (blue), loops (grey), and the locations of VR residues are indicated. The core secondary structure elements (the paired β1β5 strands, the connecting α1 and α2 helices, and the β2β3β4 sheet) of the CLec-fold are labeled. Other secondary structures may form the inserts often found in CLec-folds.

Sup. Figure 2. Mutagenic homing by the *Lp Corby* DGR. (*A*) Sequence analysis of products from the PCR-based mutagenic homing assays shown in Fig. 2C lane 2

reveals mutagenesis at VR positions corresponding to adenines in the cognate TR. (*B*) The invariant tag from Figure 2B was knocked into the genomic TR in wt *Corby* cells generating *Lp Corby* TR-GC. Cells were grown in rich media, genomic DNA extracted and probed by PCR for transfer of the tag from TR to VR. Bands (P1+P3) observed with *wt* samples contain mutagenic homing products, as shown in (*C*), while bands appearing in *ΔRT* samples are a result of template switching (detected due to the increased number of cycles) and do not represent mutagenic homing products. (*C*) Sequencing of PCR products from *Corby* TR-GC (B, P1+P3) reveals mutagenesis of nucleotide positions in VR corresponding to adenines in TR.

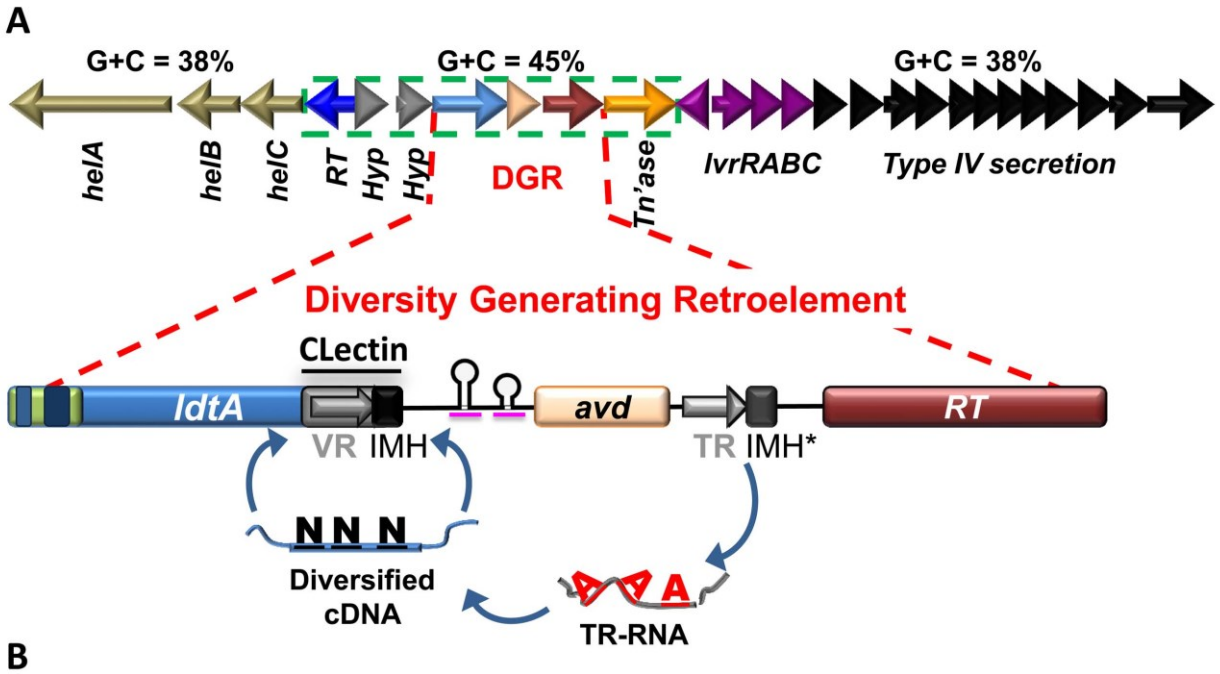
Sup. Figure 3. Detecting mutagenic homing during different growth phases of *Lp*. Various *Lp* TRwTAG strains with deletions in *relA* or *relA*/*spoT* were grown to either early exponential (A) or early stationary phase (B), total DNA was extracted and assayed for transfer of the invariant TAG from TR to VR using indicated primers. Presence (+) or absence (-) of DGR *RT*, *relA*, and *spoT* is indicated. Primers TAGFor/avdRev detected transfer of TAG to VR as a surrogate for mutagenic homing with ldtAFor/avdRev showing equal loading of DNA. Red arrows in (*B*) indicate faint bands indicative of mutagenic homing.

Sup. Figure 4. Increased rates of mutagenic homing in *relA*/*spoT* mutant cells require the DGR *RT*. An in-frame deletion of the DGR *RT* was introduced into Corby TRwTAG cells with deletions of *relA* and *spoT*. Double and triple mutant cells were

grown to ODs corresponding to growth phases: early exponential (EE), early stationary (ES), or late stationary (LS). Total DNA was extracted, treated with RNase, and assayed for the transfer of invariant TAG from chromosomal TR to VR using the following primers, TAGFor/avdRev and ldtAFor/avdRev to ensure the equal loading of DNA. Controls for detection of mutagenic homing include *relA*/*spoT* mutant cells grown to LS phase (positive) and TRwTAGΔRT (ΔRT) cells grown to late stationary phase (negative).

Sup Figure. 5. Sequencing of PCR products from *relA*/*spoT* mutant cells verifies adenine mutagenesis. PCR products amplified using primers TAGFor/avdRev on DNA extracted from TRwTAGΔ*relA*Δ*spoT* cells grown to late stationary phase, from Figure 3C, was cloned and sequenced. Comparison of several clones against predicted sequences of VR and TR containing the invariant TAG revealed adenine mutagenesis. Adenines are colored red and all other nucleotides in green. IMH and invariant TAG are indicated.

Figure 1.



B

```
VR GGC GGT TTT GCC AGT GTC TAC TAC TGG GGT TCT ACC GAG TTT TCC GGC AAT CCT ACG GTC TTC GCC TGG GAC
TR GGC GGT TTT GCC AAT GAC AAC TAC TGG AGT TCT ACC GAG TTT TCC GGC AAT CCT ACG AAC AAC GCC TGG AAC

VR CAG TTC TTC GGC GAT GGC TTC CGG TTT GGT GTC AAT AAG GTC TTC AAT CTA CGG GTT AGG TGC GTT CGG GCT
TR CAG AAC TTC AAC AAT GGC AAC CAG TTT GAT GAC AAT AAG AAC AAC AAT CTA CGG GTT AGG TGC GTT CGG GGT

VR TTT ACC CCT TAACCCCTTTATCCCTTTTGAGGCAGCGCAGCTGCCTCTTATAAAAACCGCGCAGCGGT
TR TTT AAA CAA AGC
```

Figure 2.

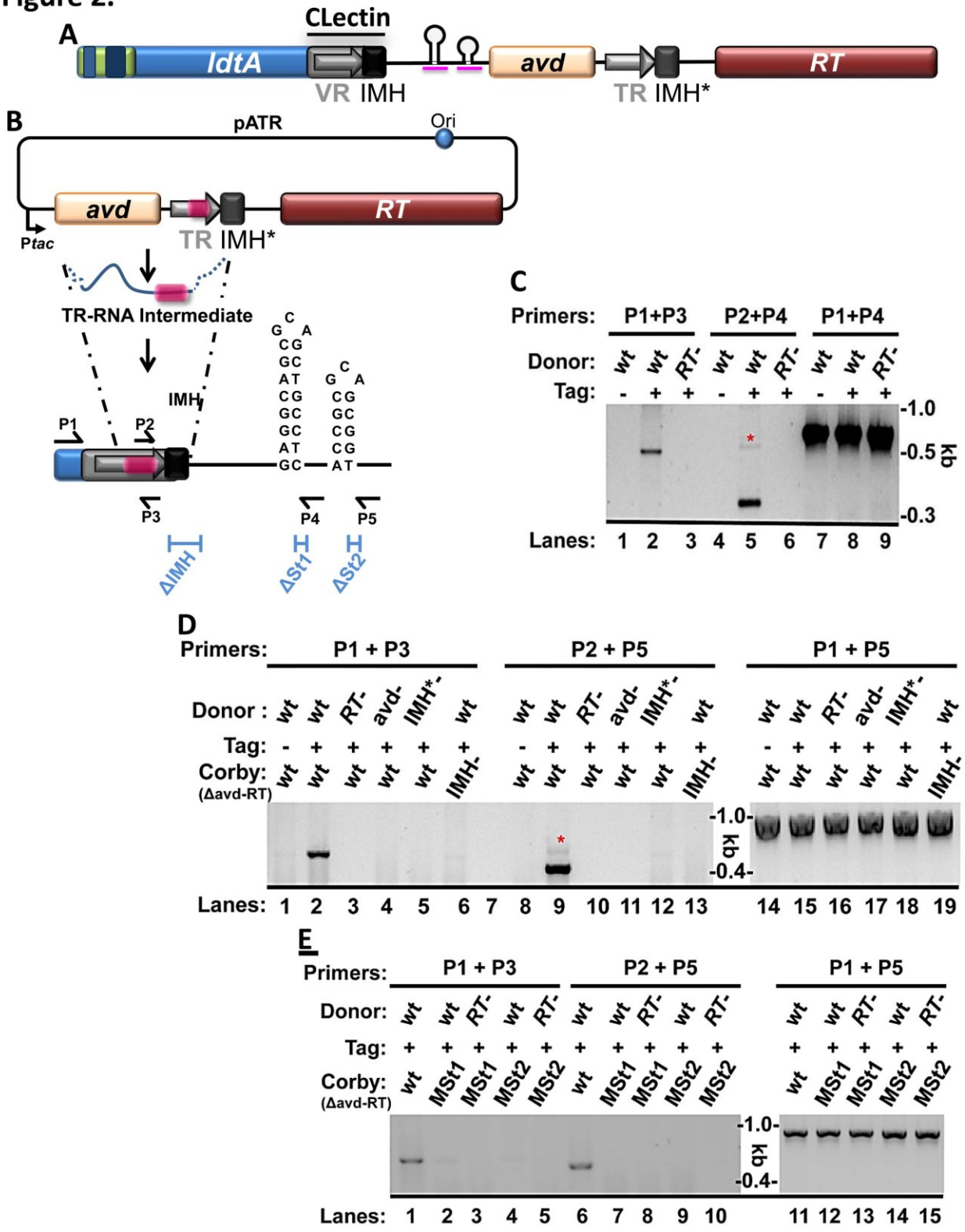
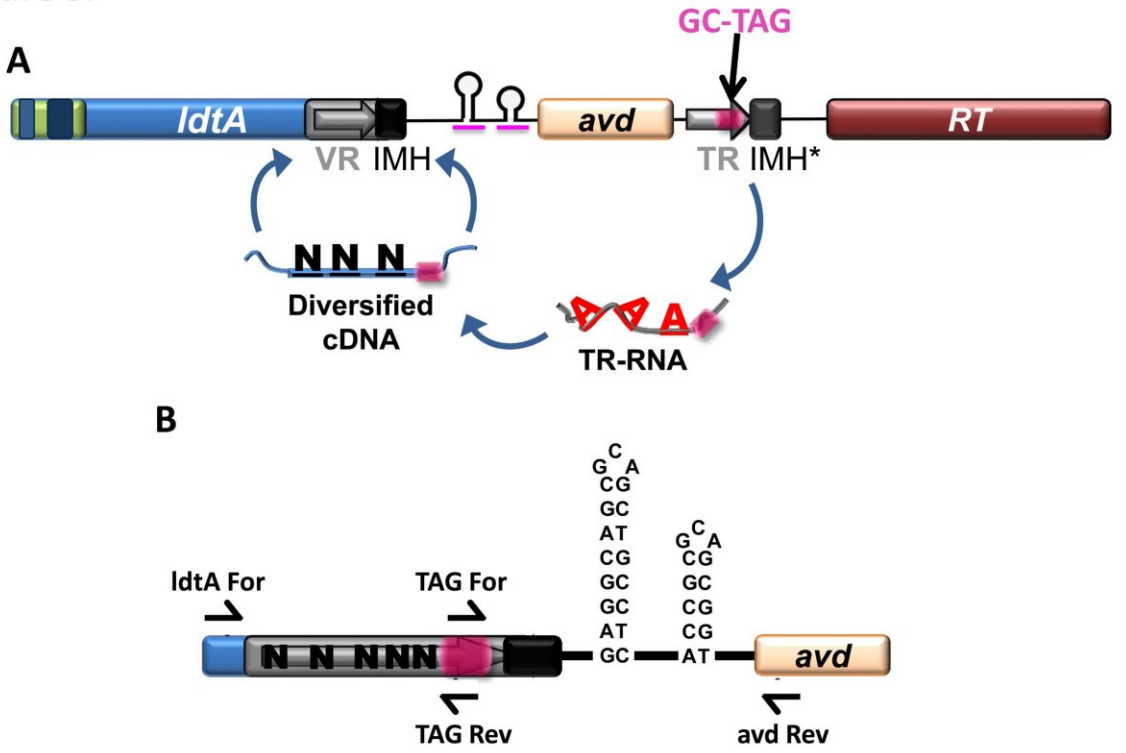

A
CLectin
IdtA
VR IMH
avd
TR IMH*
RT
B
pATR
Ori
avd
RT
Ptac
TR IMH*
TR-RNA Intermediate
IMH
P1
P2
P3
P4
P5
ΔIMH
ΔSt1
ΔSt2
C
Primers:
P1+P3
P2+P4
P1+P4
Donor:
wt
RT-
Tag:
1.0
0.5
0.3
kb
Lanes:
1 2 3 4 5 6 7 8 9
D
Primers:
P1 + P3
P2 + P5
P1 + P5
Donor :
avd-
IMH*-
Corby:
(Δavd-RT)
IMH-
1.0
0.4
Lanes:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
E
Primers:
P1 + P3
P2 + P5
P1 + P5
Donor:
Tag:
Corby:
(Δavd-RT)
MSt1
MSt2
1.0
0.4
Lanes:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Figure 3.


A
GC-TAG
ldtA
avd
RT
VR IMH
TR IMH*
N N N
Diversified cDNA
TR-RNA
B
ldtA For
TAG For
N N NNN
avd
TAG Rev
avd Rev

C

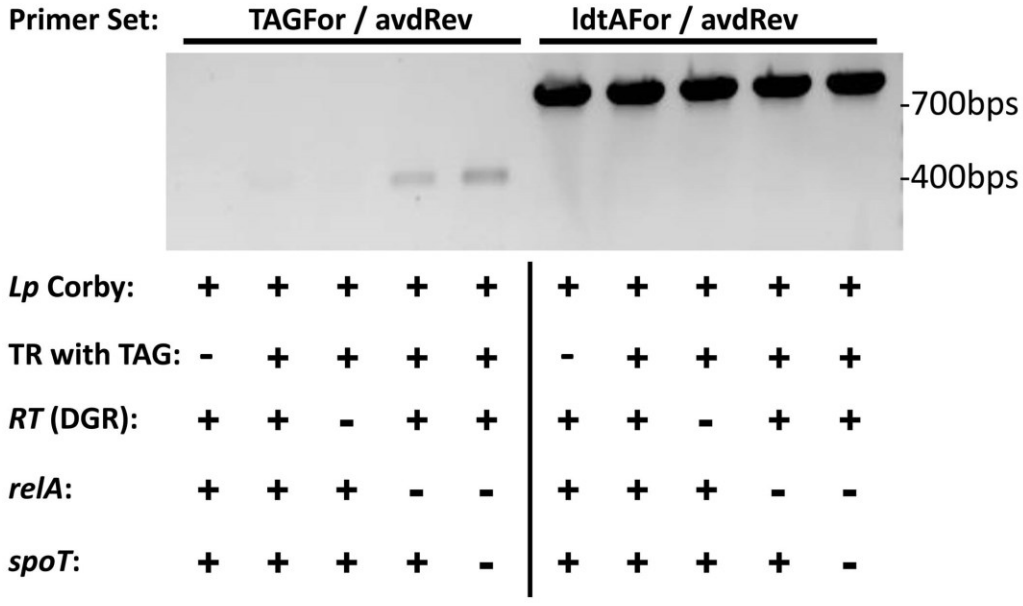

Primer Set:
TAGFor / avdRev
ldtAFor / avdRev
700bps
400bps
Lp Corby: + + + + + + + + + +
TR with TAG: - + + + + - + + + +
RT (DGR): + + - + + + + - + +
relA: + + + - - + + + - -
spoT: + + + + - + + + + -

Figure 4.

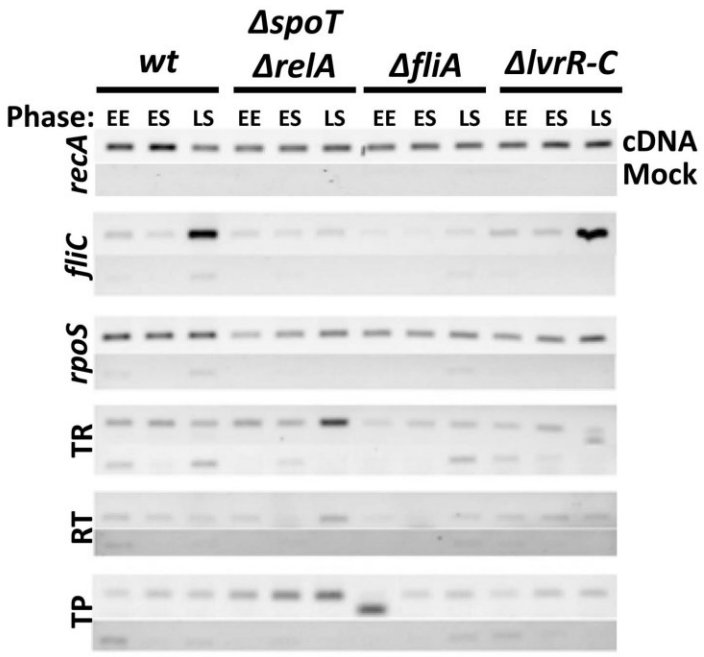

wt
ΔspoT ΔrelA
ΔfliA
ΔlvrR-C
Phase: EE ES LS EE ES LS EE ES LS EE ES LS
recA
cDNA
Mock
fliC
rpoS
TR
RT
TP

Figure 5.

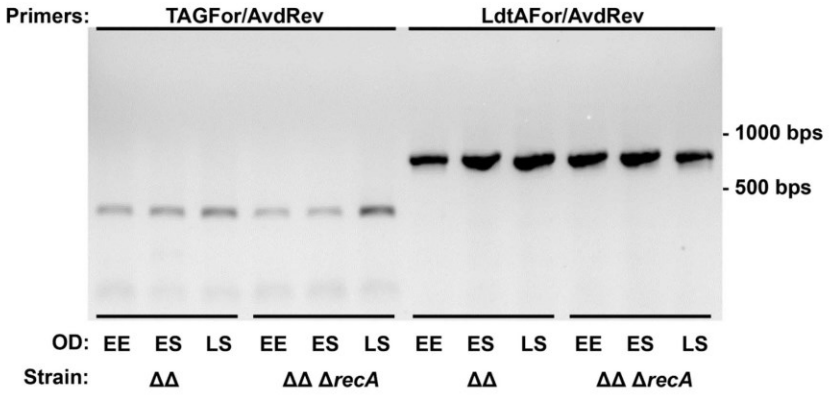

Primers:
TAGFor/AvdRev
LdtAFor/AvdRev
- 1000 bps
- 500 bps
OD: EE ES LS EE ES LS EE ES LS EE ES LS
Strain:
ΔΔ
ΔΔ Δ*recA*
ΔΔ
ΔΔ Δ*recA*

Figure 6.

A

Results

```
>gi
     36  Rnamotif - ATTAAGTCCGACCGCTGCGCGGTTTTTATAAGAG -4.50
   1292  Rnamotif - TTCAACGATTTTTCTGGCTACCTCCAGGATTAATTTCTTAT -5.70


Total number of predicted transcription terminators: 2
```

B

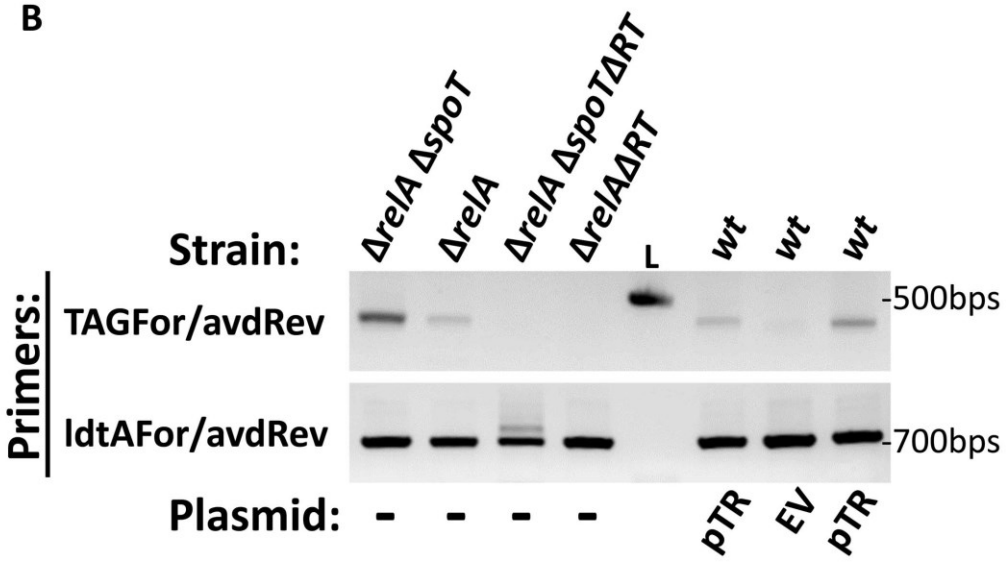
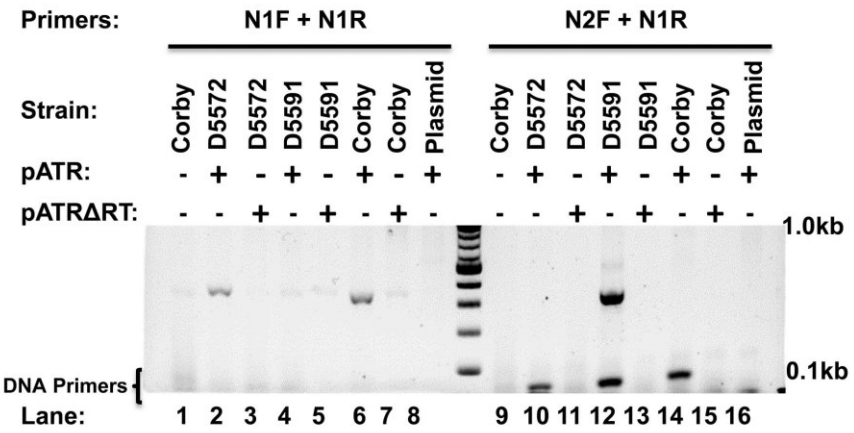
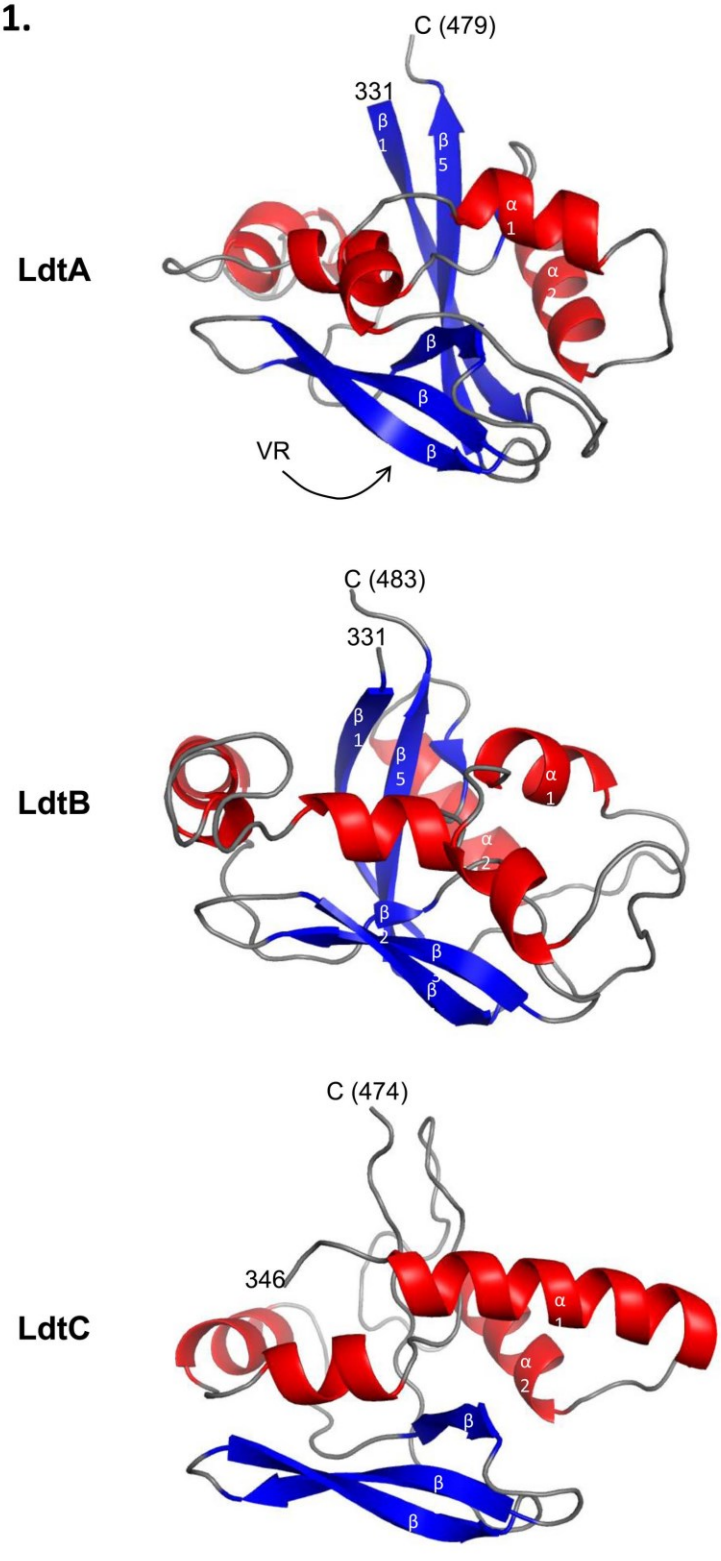


Figure 7.


Primers:
N1F + N1R
N2F + N1R
Strain:
Corby
D5572
D5572
D5591
D5591
Corby
Corby
Plasmid
Corby
D5572
D5572
D5591
D5591
Corby
Corby
Plasmid
pATR:
- + - + - + - + - + - + - + - +
pATRΔRT:
- - + - + - + - - - + - + - + -
1.0kb
0.1kb
DNA Primers
Lane:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

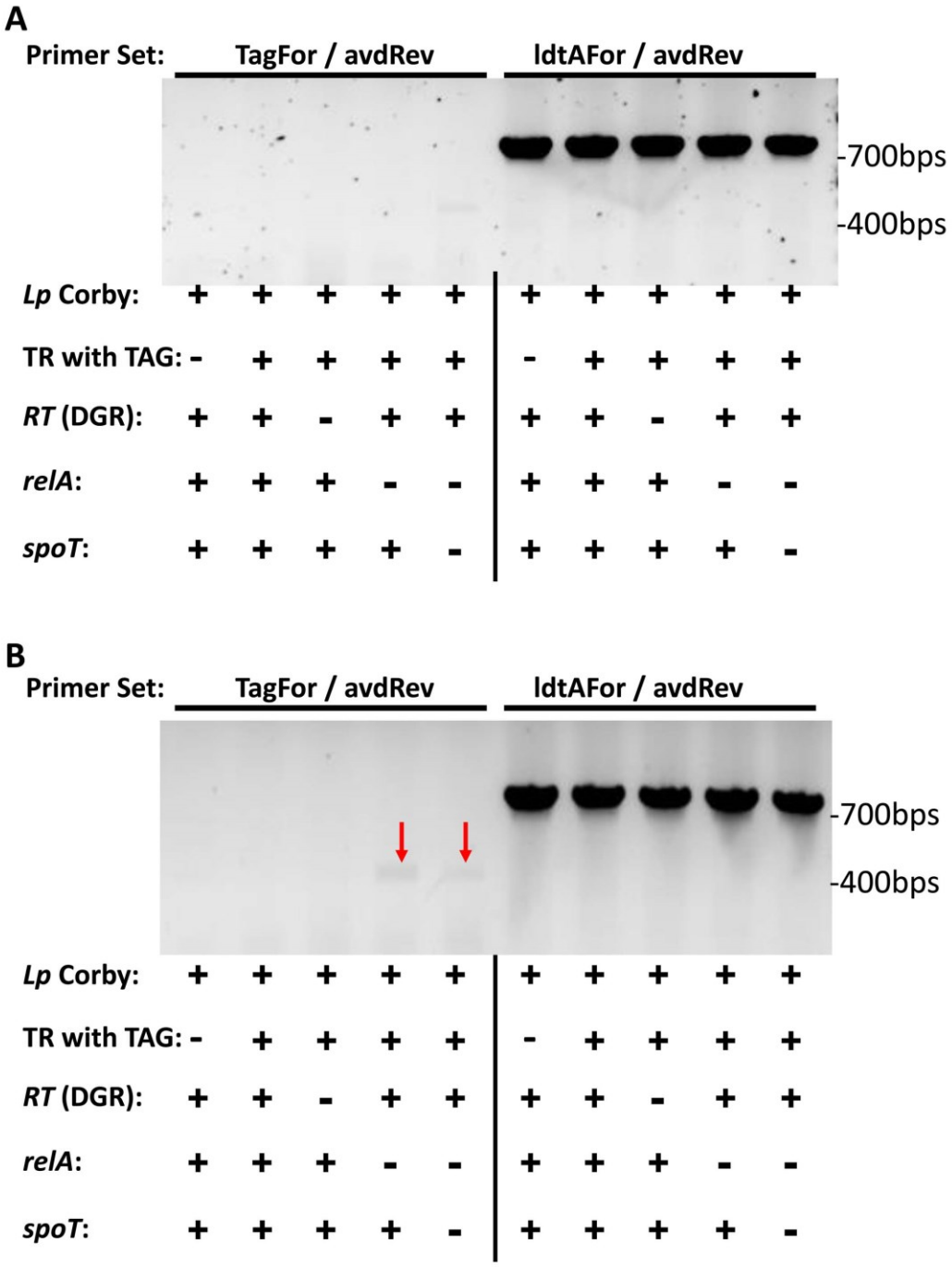
Sup. Figure 1.


LdtA
C (479)
331
β 1
β 5
α 1
α 2
β
β
β
VR
LdtB
C (483)
331
β 1
β 5
α 1
α 2
β 2
β
β
LdtC
C (474)
346
α 1
α 2
β
β
β

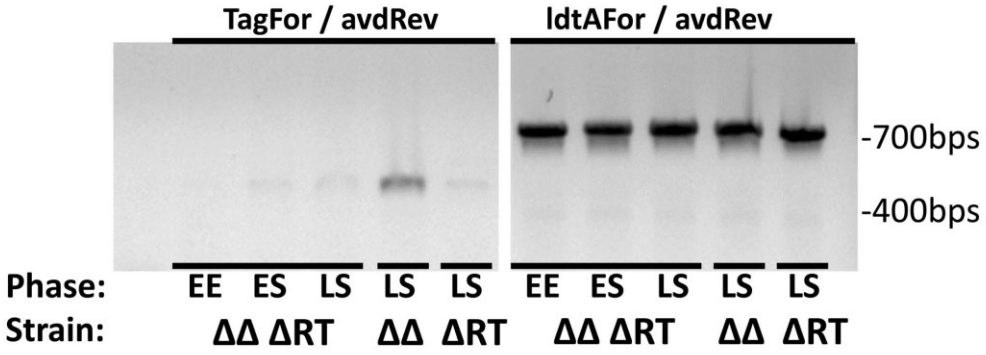
Sup. Figure 2.


A
5'
Position(nt): 1
P4
120
TR
VR
VR1
VR2
VR3
VR4
Invariant Tag
B
Primers:
P1 + P3
P2 + P5
Clone1
Clone2
Clone1
Clone2
wt
ΔRT
C
5'
Position(nt):
10
P4
129
TR
VR
VR1
VR2
VR3
VR4
Invariant Tag

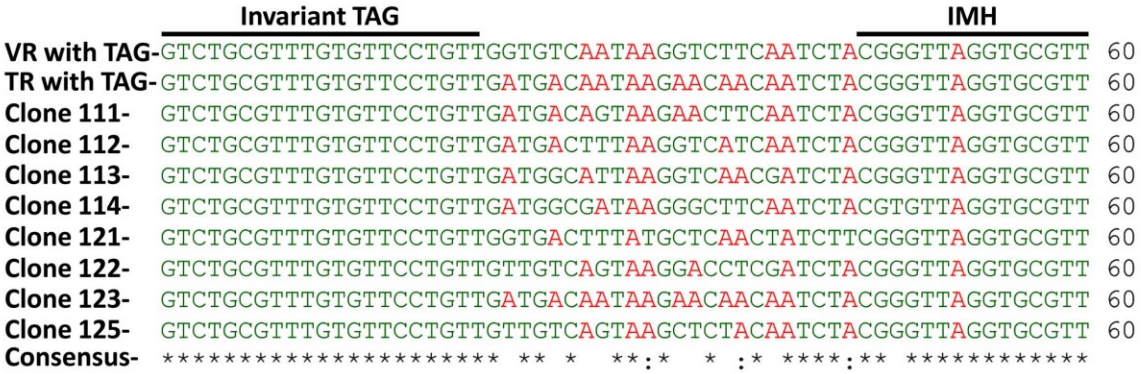
Sup. Figure 3.


A
Primer Set:
TagFor / avdRev
ldtAFor / avdRev
-700bps
-400bps
Lp Corby: + + + + + + + + + +
TR with TAG: - + + + + - + + + +
RT (DGR): + + - + + + + - + +
relA: + + + - - + + + - -
spoT: + + + + - + + + + -
B
Primer Set:
TagFor / avdRev
ldtAFor / avdRev
-700bps
-400bps
Lp Corby: + + + + + + + + + +
TR with TAG: - + + + + - + + + +
RT (DGR): + + - + + + + - + +
relA: + + + - - + + + - -
spoT: + + + + - + + + + -

Sup. Figure 4.


TagFor / avdRev
ldtAFor / avdRev
-700bps
-400bps
Phase:
EE ES LS LS LS EE ES LS LS LS
Strain:
ΔΔ ΔRT ΔΔ ΔRT ΔΔ ΔRT ΔΔ ΔRT

Sup. Figure 5.


Invariant TAG
IMH
VR with TAG-GTCTGCGTTTGTGTTCCTGTTGGTGTCAATAAGGTCTTCAATCTACGGGTTAGGTGCGTT 60
TR with TAG-GTCTGCGTTTGTGTTCCTGTTGATGACAATAAGAACAACAATCTACGGGTTAGGTGCGTT 60
Clone 111- GTCTGCGTTTGTGTTCCTGTTGATGACAGTAAGAACTTCAATCTACGGGTTAGGTGCGTT 60
Clone 112- GTCTGCGTTTGTGTTCCTGTTGATGACTTTAAGGTCATCAATCTACGGGTTAGGTGCGTT 60
Clone 113- GTCTGCGTTTGTGTTCCTGTTGATGGCATTAAGGTCAACGATCTACGGGTTAGGTGCGTT 60
Clone 114- GTCTGCGTTTGTGTTCCTGTTGATGGCGATAAGGGCTTCAATCTACGTGTTAGGTGCGTT 60
Clone 121- GTCTGCGTTTGTGTTCCTGTTGGTGACTTTATGCTCAACTATCTTCGGGTTAGGTGCGTT 60
Clone 122- GTCTGCGTTTGTGTTCCTGTTGTTGTCAGTAAGGACCTCGATCTACGGGTTAGGTGCGTT 60
Clone 123- GTCTGCGTTTGTGTTCCTGTTGATGACAATAAGAACAACAATCTACGGGTTAGGTGCGTT 60
Clone 125- GTCTGCGTTTGTGTTCCTGTTGTTGTCAGTAAGCTCTACAATCTACGGGTTAGGTGCGTT 60
Consensus- ********** ** * **:* * :* ****:** **********

References

1. Doulatov, S., et al., *Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements.* Nature, 2004. **431**(7007): p. 476-81.
2. Gogvadze, E. and A. Buzdin, *Retroelements and their impact on genome evolution and functioning.* Cell Mol Life Sci, 2009. **66**(23): p. 3727-42.
3. Liu, M., et al., *Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.* Science, 2002. **295**(5562): p. 2091-4.
4. Guo, H., et al., *Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification.* Mol Cell, 2008. **31**(6): p. 813-23.
5. Guo, H., et al., *Target site recognition by a diversity-generating retroelement.* PLoS Genet, 2011. **7**(12): p. e1002414.
6. Miller, J.L., et al., *Selective ligand recognition by a diversity-generating retroelement variable protein.* PLoS Biol, 2008. **6**(6): p. e131.
7. Schillinger, T., et al., *Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF.* BMC Genomics, 2012. **13**: p. 430.
8. data, U.
9. Shin, S., *Innate Immunity to Intracellular Pathogens: Lessons Learned from Legionella pneumophila.* Adv Appl Microbiol, 2012. **79**: p. 43-71.
10. Isberg, R.R., T.J. O'Connor, and M. Heidtman, *The Legionella pneumophila replication vacuole: making a cosy niche inside host cells.* Nat Rev Microbiol, 2009. **7**(1): p. 13-24.
11. Dalebroux, Z.D., et al., *ppGpp conjures bacterial virulence.* Microbiol Mol Biol Rev, 2010. **74**(2): p. 171-99.
12. Langille, M.G., W.W. Hsiao, and F.S. Brinkman, *Detecting genomic islands using bioinformatics approaches.* Nat Rev Microbiol, 2010. **8**(5): p. 373-82.
13. Lautner, M., et al., *Regulation, integrase-dependent excision, and horizontal transfer of genomic islands in Legionella pneumophila.* J Bacteriol, 2013. **195**(7): p. 1583-97.
14. McMahon, S.A., et al., *The C-type lectin fold as an evolutionary solution for massive sequence variation.* Nat Struct Mol Biol, 2005. **12**(10): p. 886-92.
15. De Buck, E., et al., *A putative twin-arginine translocation pathway in Legionella pneumophila.* Biochem Biophys Res Commun, 2004. **317**(2): p. 654-61.
16. Stanley, N.R., T. Palmer, and B.C. Berks, *The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in Escherichia coli.* J Biol Chem, 2000. **275**(16): p. 11591-6.
17. Chatzi, K.E., et al., *Breaking on through to the other side: protein export through the bacterial Sec system.* Biochem J, 2013. **449**(1): p. 25-37.
18. Narita, S. and H. Tokuda, *Sorting of bacterial lipoproteins to the outer membrane by the Lol system.* Methods Mol Biol, 2010. **619**: p. 117-29.
19. Shruthi, H., M.M. Babu, and K. Sankaran, *TAT-pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes.* J Mol Evol, 2010. **70**(4): p. 359-70.

20. Gralnick, J.A., et al., *Extracellular respiration of dimethyl sulfoxide by Shewanella oneidensis strain MR-1.* Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4669-74.
21. Alayyoubi, M., et al., *Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase.* Structure, 2013. **21**(2): p. 266-76.
22. Medhekar, B. and J.F. Miller, *Diversity-generating retroelements.* Curr Opin Microbiol, 2007. **10**(4): p. 388-95.
23. Dalebroux, Z.D. and M.S. Swanson, *ppGpp: magic beyond RNA polymerase.* Nat Rev Microbiol, 2012. **10**(3): p. 203-12.
24. Dalebroux, Z.D., et al., *Distinct roles of ppGpp and DksA in Legionella pneumophila differentiation.* Mol Microbiol, 2010. **76**(1): p. 200-19.
25. Al-Khodor, S., et al., *The PmrA/PmrB two-component system of Legionella pneumophila is a global regulator required for intracellular replication within macrophages and protozoa.* Infect Immun, 2009. **77**(1): p. 374-86.
26. Rasis, M. and G. Segal, *The LetA-RsmYZ-CsrA regulatory cascade, together with RpoS and PmrA, post-transcriptionally regulates stationary phase activation of Legionella pneumophila Icm/Dot effectors.* Mol Microbiol, 2009. **72**(4): p. 995-1010.
27. Molofsky, A.B., L.M. Shetron-Rama, and M.S. Swanson, *Components of the Legionella pneumophila flagellar regulon contribute to multiple virulence traits, including lysosome avoidance and macrophage death.* Infect Immun, 2005. **73**(9): p. 5720-34.
28. Gautheret, D. and A. Lambert, *Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.* J Mol Biol, 2001. **313**(5): p. 1003-11.
29. Dougan, G., et al., *The Escherichia coli gene pool.* Curr Opin Microbiol, 2001. **4**(1): p. 90-4.
30. Sidhu, S.S., et al., *Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions.* J Mol Biol, 2004. **338**(2): p. 299-310.
31. Jules, M. and C. Buchrieser, *Legionella pneumophila adaptation to intracellular life and the host response: clues from genomics and transcriptomics.* FEBS Lett, 2007. **581**(15): p. 2829-38.
32. Jepras, R.I., R.B. Fitzgeorge, and A. Baskerville, *A comparison of virulence of two strains of Legionella pneumophila based on experimental aerosol infection of guinea-pigs.* J Hyg (Lond), 1985. **95**(1): p. 29-38.
33. Ninio, S., J. Celli, and C.R. Roy, *A Legionella pneumophila effector protein encoded in a region of genomic plasticity binds to Dot/Icm-modified vacuoles.* PLoS Pathog, 2009. **5**(1): p. e1000278.
34. Kamalakkannan, S., et al., *Bacterial lipid modification of proteins for novel protein engineering applications.* Protein Eng Des Sel, 2004. **17**(10): p. 721-9.
35. Morales, V.M., A. Backman, and M. Bagdasarian, *A series of wide-host-range low-copy-number vectors that allow direct screening for recombinants.* Gene, 1991. **97**(1): p. 39-47.

Chapter 3. Distribution of Diversity Generating Retroelements within the genus *Legionella*.

Abstract

Diversity-generating retroelements (DGRs) are a unique family of retroelements which confer selective advantages to their hosts by facilitating DNA sequence evolution using a specialized error-prone reverse transcription process. Although initially identified in a *Bordetella bronchiseptica* (*Bb*) bacteriophage, BPP, comparative bioinformatic analysis of sequence databases, metagenomic datasets, and shotgun whole genome sequences identified putative DGRs in over 300 organisms which occupy a diverse range of environments and display a varied lifestyles. While DGRs are found within plasmids and phage, most are found within bacterial chromosomal elements and have been identified in all bacteria phyla with significant sequence coverage. All DGRs are structurally similar, each containing conserved elements demonstrated necessary for adenine mutagenesis, suggesting they all function through a fundamentally conserved mechanism and observed variations in architectures as well as associated components likely reflects adaptations necessary to function within a particular host. We have identified DGRs within two species and several strains of the genus *Legionella*, whose members are opportunistic human pathogens and causative agents of Legionnaires' disease and Pontiac fever. *Legionella* DGRs are found as chromosomal elements within both clinical and environmental isolates. Often they are part of horizontally acquired genomic islands which are incorporated into larger Integrative and Conjugative Element (ICE). While they encode nearly homologous DGR diversification machinery, these elements encode for a small bifurcated family of target proteins (TPs) that share C-terminal homology suggesting the variable repeat (VR) displaying C-type lectin (CLec) domain is being maintained. However, each VR displays unique patterns of adenine

mutagenesis and these TPs share little N-terminal homology. This indicates the diversification machinery is capable of functioning on a number of proteins and thus demonstrating the adaptability of DGR components. While we have identified a number of elements within the *Legionella* genus, further work is needed to determine if an associate with clade or serogroup exists and any effect on pathogenesis.

Introduction

Mobile retroelements have been proposed to play a formative role in genome evolution and, while their contribution to host fitness is under debate, they are often considered as selfish elements that rarely confer adaptive advantages to their hosts [1, 2]. The classically studied eukaryotic retroelement, human L1, is reported as the only autonomous replicating element with approximately 500,000 copies in the human genome and its mobility is often associated with a number of diseases [3]. Group II introns, a related bacterial retroelement, are found within approximately 25% of genomes and their contribution to host fitness is not entirely understood [4]. Retrons are a puzzling retroelement found in several bacterial taxa however in a particular taxon only a few species may contain retrons and, while they are still under investigation, these elements appear to be fairly innocuous to their host fitness [5]. In contrast, DGRs are found widely within the bacterial domain, are often found deeply within phyla, and are theoretically capable of generating vast amounts of nucleotide sequence variation which, in turn, generates amino acid diversity in the ligand binding domains of target proteins (TP), a massive selective advantage. In the instance of BPP, peptide diversity

is responsible for an expanded range of potential interactions allowing for adaptation to a dynamic host cell surface [6-8]. Interestingly, while non-LTR retrotransposons, group II introns, and DGRs have very different effects on host fitness, their mobility appears to function via a similar mechanism of target primed reverse transcription (TPRT) [9, 10].

DGRs generate peptide diversity by iteratively diversifying defined DNA sequences that encode the ligand biding domains of TP using a template-dependent, error prone reverse-transcriptase mediated process, termed mutagenic homing, which introduces nucleotide substitutions into a VR while preserving cis- and trans- acting elements needed for future rounds of diversification. Mutagenic homing requires a DGR-encoded reverse transcriptase (RT), an accessory variability determinant (Avd), and a template repeat (TR)-derived RNA intermediate. The TR-RNA provides a template for reverse transcription during which TR adenine residues are copied into any of the four nucleotides. The diversified cDNA displaces a VR at the 3' end of the target gene (6-7) and target recognition requires two *cis*-acting sequences, the initiation of mutagenic homing (IMH) element as well as a DNA hairpin/cruciform structure (5-7).

DGRs can be distinguished from closely related retroelements based on conserved domains within components required for mutagenic homing. The DGR encoded *RT* can be distinguished from closely related group II intron and retron reverse transcriptases through a motif that is critical for dNTP binding as well as N-terminal domains which have partial conserved similarity and may form a secondary α-helix structure that has similarity to domains involved in binding of template RNA suggesting a role in TPRT [11]. Interestingly, many TRs overlap with the 5' of *RT* ORF and the BPP

TR-RNA has been shown to include part of *brt*, the BPP DGR RT (H. Guo personal communication). The C-terminus of DGR RTs shown more variation in sequence and, while they are predicted to form α-helices similar to group II intron RTs, a specific function has not yet been described (M. Gingery personal communication). DGRs are also identified by the presence of cognate TR/VR pairs within DNA regions flanking the *RT*, with the VR being found in the 3' or middle of a TP encoding open reading frame [7, 12]. While shown to be necessary for mutagenic homing, the identification of genes encoding DGR accessory proteins is difficult because they can share little sequence homology or, in the case of *Bacteroides* elements, similar DGRs can contain completely different accessory proteins. The identification of genetic factors like DGR stem loop/cruciform structures as well as IMH/IMH* is also often difficult as they appear to only be conserved on the species level.

Analysis of nucleotide databases, deposited whole shotgun sequences, and metagenomic datasets using custom scripts or alignments to identified DGR specific components and genes identified hundreds of putative elements [8, 12, 13]. To date DGRs have been identified in ~300 organisms found in at least 20 phyla of bacteria and one phyla of archaea (B. Paul personal communication and unpublished data). These organisms occupy a wide number of ecological niches which range from terrestrial to marine as well as from microaerobic to arctic permafrost and display a variety of lifestyles from commensal, planktonic, free living, or pathogenic [8]. Bacterial DGRs appear to be enriched in three phyla of bacteria, *Bacteroidestes*, *Firmicutes*, and *Proteobacteria* and recent work has identified a vast number of DGRs associated with phage that occupy the human intestinal gut [12]. The human oral pathogen *Treponema*

denticola contains an element that can diversify up to seven TPs which are likely found upon the cell surface. At least one of these TPs has partial similarity to the human formylglycine-generating enzyme (hFGE) suggesting it might have enzymatic activity but it has been suggested that the other TPs may play a role in *T. denticola*'s adhesion to individuals of a microbial community as it attempts to colonize the subgingival plaque [14]. *Trichodesmium erythraeum* is a free-living filamentous marine cyanobacteria shown to contribute to nitrogen fixation [15, 16]. It has two DGRs predicted to diversify 14 TPs and many of these proteins have N-terminal regions with homology to known serine/threonine protein kinase (STPK) domains (M. Gingery personal communication) [8]. This suggest these proteins are being post-translationally modified by host machinery and are playing a role in signal transduction pathways that could be modulating host physiological responses. Finally, putative DGRs were identified in sequences within the human gut virome as well as metagenomic datasets which contain homology to the BPP DGR phage tail binding gene *mtd* [12]. This suggests the genes encoded by these DGRs, like BPP, function in expanding the range of host ligands. Cumulatively, the identified DGRs could diversify hundreds of TPs, each with a potentially different biological function.

In silico analysis identified a putative DGR within a sequenced strain of the opportunistic human pathogen *Legionella pneumophila* (*Lp*), [17]. *Lp* is comprised of over 64 serogroups but more than 84% of disease is caused by members of serogroup I (Sg1) [18, 19], which contains both clinical and environmental isolates that cause disease and have a high degree of genome plasticity [20, 21]. The *Lp* DGR is found within a hyper-virulent clinical isolate called Corby [22]. It encodes all components

characteristically shared amongst DGRs and is an active element in that it is capable of directing adenine-specific mutagenesis of a TP encoding gene, *ldtA*. While DGRs were only identified in one sequenced *Lp* Sg1 strains, we were curious to determine if these elements were in other *Legionella* strains and species. Using a polymerase chain reaction (PCR) based assay to identify conserved components of DGR, we screened a library of *Lp* strains and identified elements that are highly homologous. All *Lp* DGRs maintain highly similar diversification machinery but have co-opted different carrier sequences to display related variable domains demonstrating the modularity of these systems. We expanded our query to deposited sequence databases and identified a putative element within another species of *Legionella, L. tunisiensis* [23]. We have combined a library screen with *in silico* analysis databases to explore the distribution of DGRs within the genus *Legionella* and have identified an active retroelement that appears to have been widely distributed.

Results

The distribution of DGRs in *Lp* isolates. The DGR in *Lp* Corby, a Sg1 clinical isolate, is absent in several other sequenced Sg1 strains and conserved flanking sequences define the endpoints of the DGR genomic island (Figure 1). To further investigate the distribution of DGRs within *Lp* we conducted a PCR screen of 12 additional Sg1 clinical isolates and found that three contained DGR-associated genes (Sup. Figure 1A, B). To determine their composition and architectures, multiplexed genomic libraries of the newly identified DGR-containing isolates were sequenced.

Complete sets of conserved DGR components, including *RT*, *avd*, TR, VR, stem loops, IMH, IMH*, and TP loci were identified in strains D5572 and D5591. The third isolate, D5549, contained *RT*, TR and *avd* genes without an identifiable linked TP. However, given the ability of the diversifying machinery to act *in trans* [10], unlinked target proteins may exist in this isolate. In D5591, the DGR is located on a genomic island similar to the one in *Lp* Corby, with identical regulatory genes to the right (Figure 1, Sup. Figure 1D). A chromosomal rearrangement at the left boundary appears to have substituted different flanking loci and removed genes from the genomic island. For D5572 and D5549, DGRs are encoded on similar genomic islands but flanking sequences could not be identified, even at 50-fold coverage, as the high degree of plasticity in the *Lp* genomes precluded assembly of long stretches of contiguous sequence (Sup. Figure 1C) [24].

Figure 2 shows details of the DGRs of three Sg1 isolates and their comparison is quite interesting. *RT* and *avd* genes are predicted to encode nearly identical products and the TR loci share 97% nucleotide identity (Sup. Figure 2A). VR sequences are also similar with differences at positions corresponding to adenine residues in their cognate TRs (Sup. Figure 2B). Protein threading predicts that VR domains of LdtA, LdtB, and LdtC adopt similar CLec folds with binding pockets composed of diversified residues (Sup. Figure 3). Furthermore, structure based sequence alignments of several known and predicted TP CLec domains identified a conserved “GGxW” (where x is any amino acid) motif [25]. Similarly, all *Lp* TPs display CLec domains with a conserved motif which has an insert, resulting in “GGxAxxYW”. While these conserved motifs are found at the 5’ of VRs and are thought to stabilize the CLec fold, their exact structural or

biological function is unknown [14, 25]. While LdtA and LdtB are related throughout their entire length, including secretion and localization signals which predict their cell surface localization, sequences upstream of the LdtC VR have significantly diverged. Despite this, the N-terminus of LdtC is predicted to encode a Sec-dependent secretion signal followed by a lipobox motif which could provide an alternative means for surface localization [26]. This illustrates the modular nature of DGRs and is consistent with the notion that *trans*-acting factors and diversified scaffolds act in a generic manner to evolve ligand binding specificities which can be adapted to different functions through connection with different N-terminal domains.

Distribution of DGRs in deposited sequence databases. Recently, several projects to sequence clinical as well as environmental isolates of the genus *Legionella* have been undertaken with results being deposited into publically available repositories [19]. Using the Corby TR and *RT* as a signature, we analyzed these databases for putative DGRs within various species of *Legionella* and identified at least three putative DGRs. Two of these elements are found in different strains of *Lp* and the third within a close relative, *Legionella tunisiensis* (*Lt*). The draft genome of *Lt* strain LegM consists of 13 scaffolds which reveal a genome of approximately 3.5 Mbps which is predicted to have a G+C content that is similar to other *Legionella* species of 39% and analysis of 16S ribosomal RNA gene identified *Legionella feeleii* as its closest relative [23]. The *Lt* DGR is predicted to be found within a similar genomic island as the *Lp* DGRs as it is flanked by genes homologous to those found in the Corby genomic island including a predicted transposase (Figure 1). This genomic island has a G+C content of 45% as compared to an average of 39% for the rest of the genome. The *Lt* DGR TR is 148 bps

long with 96% nucleotide homology to the Corby TR and contains 43 adenines (Sup. Figure 2A). The *RT* and *avd* are predicted to encode proteins that share roughly 90% amino acid homology with components found in Corby. The *Lt* DGR contains one DNA stem loop/cruciform structure that is similar to one of two elements which were found to be essential for mutagenic homing in Corby. Interestingly, in the *Lt* element the stem is identical to the Corby element but the loop sequence has replaced “GCA” with “TAG” and alterations of the loop sequences have been shown to negatively affect levels of mutagenic homing in *Bb* BPP (Sup. Figure 4) [10]. The distance between the stem loop/cruciform structure and the IMH is conserved across all *Legionella* DGRs which is consistent with its importance in mutagenic homing [10]. The *Lt* DGR TP shares 88% amino acid homology with the D5591 DGR TP, LdtC, and maintains both the N-terminal signal peptide as well as the C-terminal CLec conserved amino acid motif “GGxAxxYW” which is likely necessary to display diversity. Furthermore, comparison of the VRs shows that while they are 84% identical they each have specific patterns of adenine mutagenesis (Sup. Figure 2B, C).

Discussion

To date, over 300 unique DGRs have been identified in phage, plasmid, or bacterial genomes and are associated with an array of diverse ecological niches. Despite their widespread distribution in nature and capacity to confer selective advantages, only a single, phage-encoded DGR has been studied in mechanistic detail

[6, 9, 27, 28]. Our discovery of a functional DGR in *Lp* which is capable of mutagenic homing provides a bacterial system for comparative analysis.

We identified retroelements highly similar to the DGR found in *Lp* strain Corby in several other *Lp* Sg1 isolates and in a close relative *Lt*. In our initial screen for DGR elements in *Lp* identified *RT* and *avd* homologs in 25% of the strains tested. Putative DGRs were identified within clinical *Lp* isolates as well as an environmental isolate from a hypersaline lake, *Lt* strain LegM [23]. It will be important to investigate the distribution of DGRs in larger sample sets which include both clinical and environmental isolates to determine if correlations exist between the presence or absence of DGRs, or the nature of variable proteins, and virulence for humans.

Our analysis revealed several interesting observations about these putative DGRs and their relationship to the element described in Corby. The *Lp* Corby, D5591, and *Lt* LegM DGRs are located in conserved genomic islands that bear the hallmarks of recent horizontal acquisition. In *Lp* Corby, and likely D5591, the DGR island appears to be a recently acquired element within a much larger ICE which appears, on the basis of G+C content, to be ancestral within *Legionellaceae* [29]. Interestingly, the Corby DGR containing ICE can be horizontally transferred between *Lp* strains as well as between *Legionella* species [29]. While an ICE could offer a means of dissemination throughout bacterial populations, DGRs do not seem to be restricted to them as the element in *Lt* is not found within an identifiable ICE. Our analysis revealed that all *Lp* DGRs share nearly identical diversification machinery: *avd*, TR, and *RT* loci as well as the targeting elements IMH and stem loops/cruciform structure. Analysis of the stem loop/cruciform

structure is interesting as the stem is conserved but the loop shows variation between strains and species. It has been reported that alteration of the *Bb* stem loop/cruciform structure, specifically in the loop sequence, diminished mutagenic homing efficiency [10]. It is currently unknown if these nucleotide substitutions have an effect on mutagenic homing or if these substitutions represent adaptations to host specific machinery. Finally, we have identified genes which are conserved between *Legionella* DGR containing genomic island. Some genes, like the transposase, have predicted functions which would explain their retention. However, several small open reading frames are also being maintained and their contribution to DGR mutagenic homing or mobility of the genomic island is unknown.

We have demonstrated that all *Lp* DGRs contain similar diversification machinery, that each strain is capable of supporting mutagenic homing (Chapter 2), and that each VR displays a unique pattern of adenine-specific mutagenesis. This is consistent with the hypothesis that these elements are active in nature. The observation that similar VRs have been fused to entirely different N-terminal sequences in *Lp* D5591/*Lt* LegM vs*. Lp* Corby/*Lp* D5572 provides further illustration of the modular nature of diversified proteins and the versatility of the VR-encoded CLec scaffold.

Materials and methods

Bacterial strains. All bacterial strains used in this chapter were a kind gift from Dr. Natalia Kozak, Centers for Disease Control and Prevention.

PCR Screen for DGR Genes. To screen *Lp* strains for putative DGR genes a series of degenerative primers were designed using previously described methods [30]. The first iteration of primers used is as follows: 16SFor- AGCATKGTCTAGCTTGCTAG, 16SRev- TCCTCCCCACTGAAAGTG, avdFor- TGTTTGAGGTAACGAAAGATTTC, avdRev- CCGGTCACCTGCTTGCCTA, RTFor- AAATCATCGACGTAACGACCATA, RTRev- CTTTCGTGACCGTGTGGTGC.

Sequencing of *Lp* Strains. The genomes of *Lp* strains D5549, D5572, and D5591 were assembled into multiplexed libraries and sequenced on an Illumina platform as previously described [31]. Reads were assembled into contigs using program Assembler and nucleotide sequences of genomic regions were visualized for analysis using Artemis software from the Sanger Center [32].

Sequence Analysis Tools. Bacterial genome, nucleotide, and protein sequence data were obtained from the National Center for Biotechnology Information (NCBI) databases [19]. Sequence comparisons were performed by NCBI BLAST [33] or by use of Vector NTI (Life Technologies), while multiple sequence alignments were performed using ClustalW [34]. Nucleotide sequence analysis to identify nucleotide features was performed using the program REPuter to identify DNA repeat regions [35], DNA stem loop/cruciform structure was identified using Mfold [36]. Predicted subcellular localization of target proteins was performed using PSORT v.3.0 [37], lipoproteins were predicted using LipoP [38], while SEC and TAT protein trafficking motifs were predicted using SignalP [26] or TatP [39], respectively.

Figure Legends

Figure 1. A subset of *Legionella* strains contain DGRs. The genomic island (dashed green box) containing the Corby DGR is absent in sequenced and assembled genomes of *Lp* strains Philadelphia and Lens (data not shown), while DGR and flanking genomic island sequences are present in Sg1 strains D5549, D5572, and D5591, which were partially sequenced in this study as well as within another species, *Lt* strain LegM. Two hypothetical proteins (dark grey) and a gene annotated as a transposase (orange) are also conserved within the genomic island. Outside the genomic island, genes annotated as hypothetical (light grey), with homology to genes found on a plasmid within *Lp* Lens (dark blue), to homology to genes encodes in *Legionella oakridgensis* (green), heavy metal transport system (brown), or genes with homology to T4SS regulators (purple) are indicated.

Figure 2. Comparison of Legionella DGRs. Graphical representation of sequence comparisons between DGRs and the two conserved hypothetical proteins. TP genes (*ldtA*-*D*), secretion signals (green and grey), and other DGR components are shown with percent identities between nucleotide (black) or amino acid sequences (red). Predictive programs (*Material and Methods*) identified similar stem loops (St1/2) in Corby and D5572, while strain D5591 contains a single, highly structured stem loop (St3). The predicted CLec folds of LdtA, LdtB and LdtC are modeled in Sup. Figure 3.

Sup. Figure 1. A subset of *Lp* strains contain DGRs. (*A*) PCR screen to detect conserved nucleotide motifs in *avd*, *RT*, and *Lp* specific 16S rRNA using *Lp* strain 130b

(DGR-) and strain Corby (DGR+) as controls. (*B*) A library of clinical *Lp* isolates was screened for putative DGR genes using primers from (*A*). Results from screens for conserved *RT* genes and 16S loci are shown. Primers against *avd* gave similar results as those for *RT*. (*C*) Genomic libraries of putative DGR-containing strains were generated and sequenced. Reads for D5591 were assembled into contigs and a pairwise assembly (*Material and Methods*) using Corby as a reference genome is shown. (*D*) The average G+C content was determined using 200 segments for a 10 kbp region containing the putative DGR island (dashed green box) for strain D5591. The average G+C content for the entire island is 46% while the average G+C content for the flanking regions is 35% (dashed red lines).

Sup. Figure 2. Analysis of *Lp* TRs and VRs. Alignments of TRs and VRs from *Lp* Corby, D5572, and D5591 as well as *Lt* LegM showing nucleotide differences from the Corby sequences. Alignment of (*A*) TRs (*B*) VRs and (*C*) Corby TR compared against VRs demonstrate highly conserved TRs and strain specific adenine mutagenesis.

Sup. Figure 3. Predicted structures of the C-terminal domains of LdtA, LdtB, and LdtC in ribbon representation. α-helices (red), β-strands (blue), loops (gray), and the locations of VR residues are indicated. The core secondary structure elements (the paired β1β5 strands, the connecting α1 and α2 helices, and the β2β3β4 sheet) of the CLec-fold are labeled. Other secondary structures may form the inserts often found in CLec-folds.

Sup. Figure 4. Conservation of DGR stem loops. Graphical DNA motif demonstrating conservation of *TP* and stem loop/cruciform structure for all *Legionella* DGRs depicted in Figure 2. *TP* stop codon, 8 nt stem, and 3 nt loop sequences are indicated.

Figure 1.

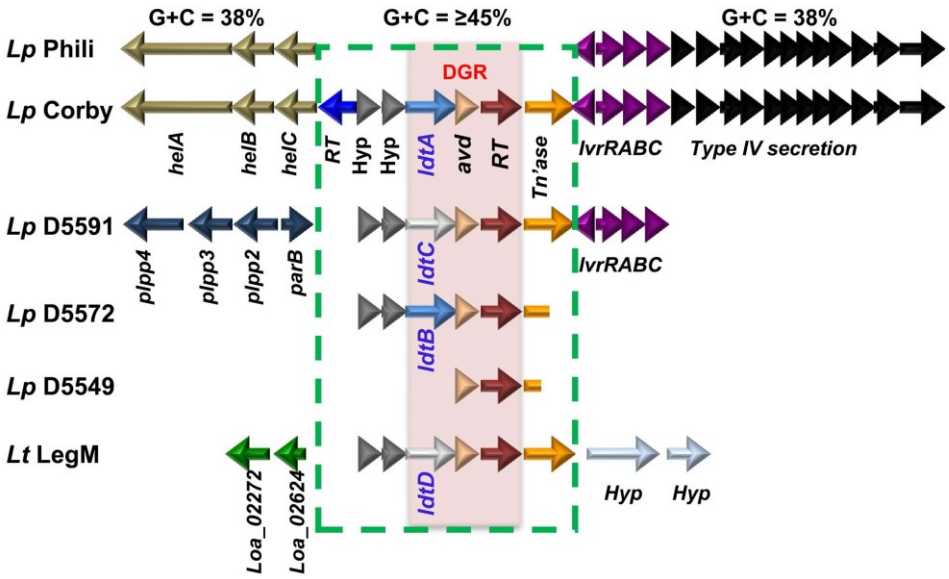
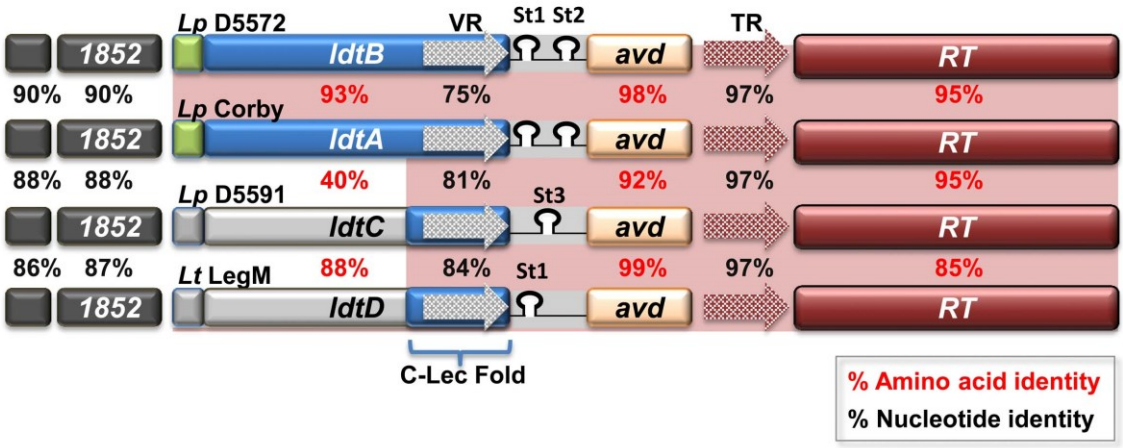
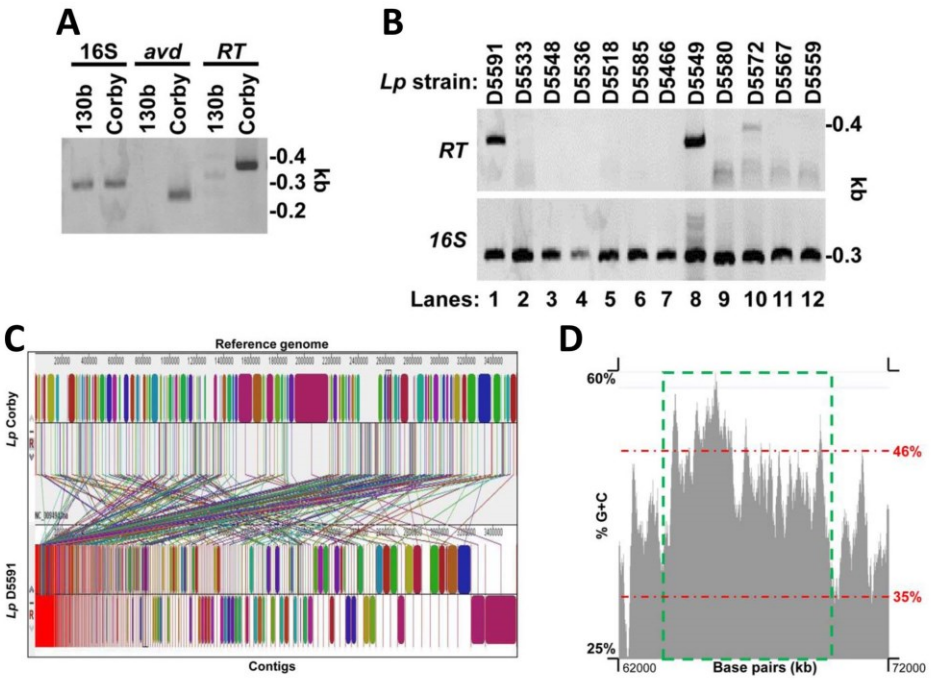

G+C = 38%
G+C = ≥45%
G+C = 38%
Lp Phili
Lp Corby
Lp D5591
Lp D5572
Lp D5549
Lt LegM
DGR
helA
helB
helC
RT
Hyp
Hyp
ldtA
avd
RT
Tn'ase
lvrRABC
Type IV secretion
plpp4
plpp3
plpp2
parB
ldtC
lvrRABC
ldtB
ldtD
Loa_02272
Loa_02624
Hyp
Hyp

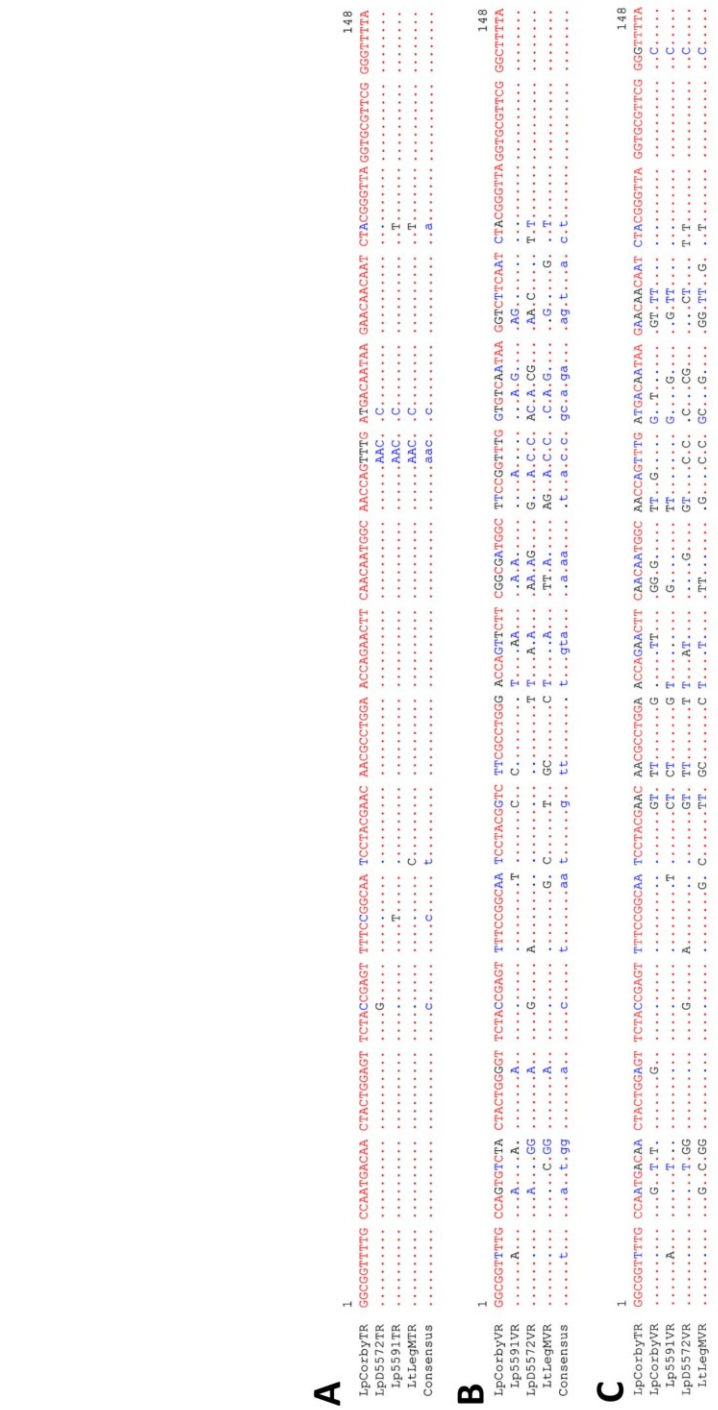
Figure 2.


Lp D5572
VR
St1 St2
TR
1852
IdtB
avd
RT
90%
90%
93%
75%
98%
97%
95%
Lp Corby
1852
IdtA
avd
RT
88%
88%
40%
81%
92%
97%
95%
Lp D5591
St3
1852
IdtC
avd
RT
86%
87%
Lt LegM
88%
84%
St1
99%
97%
85%
1852
IdtD
avd
RT
C-Lec Fold
% Amino acid identity
% Nucleotide identity

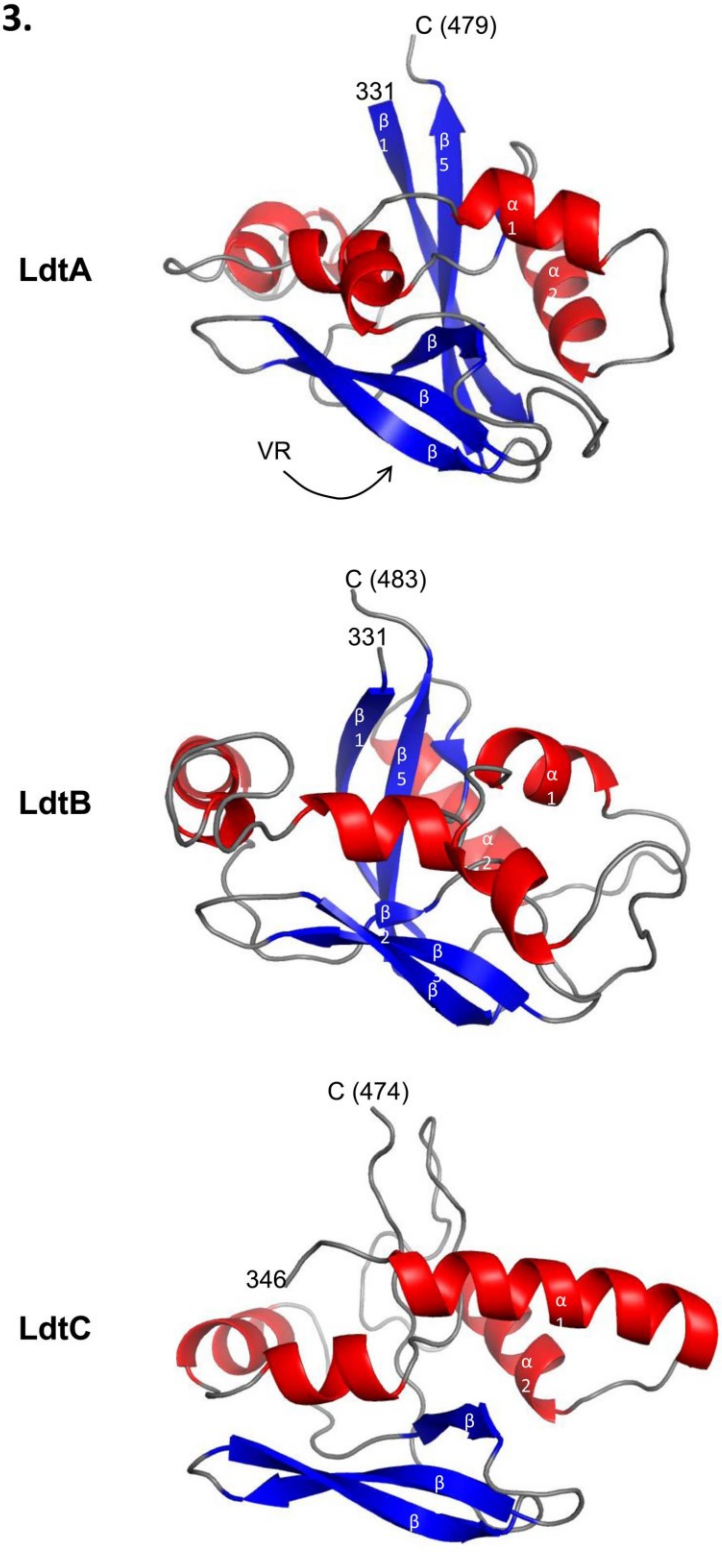
Sup. Figure 1.


A
16S
avd
RT
130b
Corby
130b
Corby
130b
Corby
-0.4
-0.3
-0.2
kb
B
Lp strain:
D5591
D5533
D5548
D5536
D5518
D5585
D5466
D5549
D5580
D5572
D5567
D5559
RT
16S
-0.4
-0.3
kb
Lanes: 1 2 3 4 5 6 7 8 9 10 11 12
C
Reference genome
Lp Corby
Lp D5591
Contigs
D
60%
46%
35%
25%
% G+C
62000
Base pairs (kb)
72000

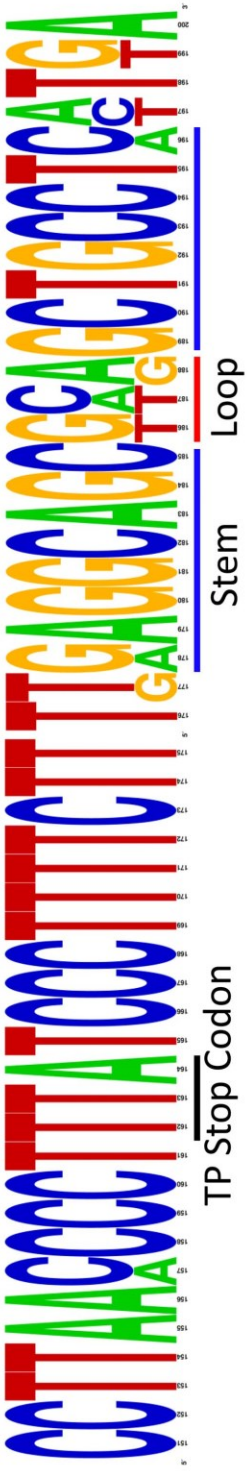
Sup. Figure 2.



Sup. Figure 3.


LdtA
C (479)
331
β 1
β 5
α 1
α 2
β
β
β
VR
LdtB
C (483)
331
β 1
β 5
α 1
α 2
β 2
β
β
LdtC
C (474)
346
α 1
α 2
β
β
β

Sup. Figure 4.



References

1. Gogvadze, E. and A. Buzdin, *Retroelements and their impact on genome evolution and functioning.* Cell Mol Life Sci, 2009. **66**(23): p. 3727-42.
2. Gordo, I., L. Perfeito, and A. Sousa, *Fitness effects of mutations in bacteria.* J Mol Microbiol Biotechnol, 2011. **21**(1-2): p. 20-35.
3. Carreira, P.E., S.R. Richardson, and G.J. Faulkner, *L1 retrotransposons, cancer stem cells and oncogenesis.* FEBS J, 2014. **281**(1): p. 63-73.
4. Lambowitz, A.M. and S. Zimmerly, *Group II introns: mobile ribozymes that invade DNA.* Cold Spring Harb Perspect Biol, 2011. **3**(8): p. a003616.
5. Lampson, B.C., S. Inouye M Fau - Inouye, and S. Inouye, *Retrons, msDNA, and the bacterial genome.* (1424-859X (Electronic)).
6. Doulatov, S., et al., *Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements.* Nature, 2004. **431**(7007): p. 476-81.
7. Liu, M., et al., *Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.* Science, 2002. **295**(5562): p. 2091-4.
8. Medhekar, B. and J.F. Miller, *Diversity-generating retroelements.* Curr Opin Microbiol, 2007. **10**(4): p. 388-95.
9. Guo, H., et al., *Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification.* Mol Cell, 2008. **31**(6): p. 813-23.
10. Guo, H., et al., *Target site recognition by a diversity-generating retroelement.* PLoS Genet, 2011. **7**(12): p. e1002414.
11. Blocker, F.J., et al., *Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase.* RNA, 2005. **11**(1): p. 14-28.
12. Minot, S., et al., *Hypervariable loci in the human gut virome.* Proc Natl Acad Sci U S A, 2012. **109**(10): p. 3962-6.
13. Schillinger, T., et al., *Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF.* BMC Genomics, 2012. **13**: p. 430.
14. Le Coq, J. and P. Ghosh, *Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement.* Proc Natl Acad Sci U S A, 2011. **108**(35): p. 14649-53.
15. Brandes, J.A., A.H. Devol, and C. Deutsch, *New developments in the marine nitrogen cycle.* Chem Rev, 2007. **107**(2): p. 577-89.
16. Deutsch, C., et al., *Spatial coupling of nitrogen inputs and losses in the ocean.* Nature, 2007. **445**(7124): p. 163-7.
17. Fields, B.S., R.F. Benson, and R.E. Besser, *Legionella and Legionnaires' disease: 25 years of investigation.* Clin Microbiol Rev, 2002. **15**(3): p. 506-26.
18. Kozak, N.A., et al., *Distribution of lag-1 alleles and sequence-based types among Legionella pneumophila serogroup 1 clinical and environmental isolates in the United States.* J Clin Microbiol, 2009. **47**(8): p. 2525-35.

19. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2010. **38**(Database issue): p. D5-16.
20. Lomma, M., et al., *Legionella pneumophila - Host Interactions: Insights Gained from Comparative Genomics and Cell Biology.* Genome Dyn, 2009. **6**: p. 170-86.
21. Gomez-Valero, L., C. Rusniok, and C. Buchrieser, *Legionella pneumophila: population genetics, phylogeny and genomics.* Infect Genet Evol, 2009. **9**(5): p. 727-39.
22. Jepras, R.I., R.B. Fitzgeorge, and A. Baskerville, *A comparison of virulence of two strains of Legionella pneumophila based on experimental aerosol infection of guinea-pigs.* J Hyg (Lond), 1985. **95**(1): p. 29-38.
23. Pagnier, I., et al., *Genome sequence of Legionella tunisiensis strain LegM(T), a new Legionella species isolated from hypersaline lake water.* J Bacteriol, 2012. **194**(21): p. 5978.
24. Cazalet, C., et al., *Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity.* Nat Genet, 2004. **36**(11): p. 1165-73.
25. McMahon, S.A., et al., *The C-type lectin fold as an evolutionary solution for massive sequence variation.* Nat Struct Mol Biol, 2005. **12**(10): p. 886-92.
26. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
27. Alayyoubi, M., et al., *Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase.* Structure, 2013. **21**(2): p. 266-76.
28. Miller, J.L., et al., *Selective ligand recognition by a diversity-generating retroelement variable protein.* PLoS Biol, 2008. **6**(6): p. e131.
29. Lautner, M., et al., *Regulation, integrase-dependent excision, and horizontal transfer of genomic islands in Legionella pneumophila.* J Bacteriol, 2013. **195**(7): p. 1583-97.
30. Linhart, C. and R. Shamir, *The degenerate primer design problem: theory and applications.* J Comput Biol, 2005. **12**(4): p. 431-56.
31. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
32. Rutherford, K., et al., *Artemis: sequence visualization and annotation.* Bioinformatics, 2000. **16**(10): p. 944-5.
33. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
34. McWilliam, H., et al., *Analysis Tool Web Services from the EMBL-EBI.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W597-600.
35. Kurtz, S., et al., *REPuter: the manifold applications of repeat analysis on a genomic scale.* Nucleic Acids Res, 2001. **29**(22): p. 4633-42.
36. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.
37. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.

38. Juncker, A.S., et al., *Prediction of lipoprotein signal peptides in Gram-negative bacteria.* Protein Sci, 2003. **12**(8): p. 1652-62.
39. Bendtsen, J.D., et al., *Prediction of twin-arginine signal peptides.* BMC Bioinformatics, 2005. **6**: p. 167.

Chapter 4. Analysis of the *Lp* surface displayed target proteins.

Abstract

Diversity-generating retroelements (DGRs) are a family of retroelements that are able to iteratively diversify defined DNA sequences that encode the ligand biding domains of target proteins (TP). Diversification occurs through a template-dependent, error prone reverse-transcriptase mediated process, termed mutagenic homing, which introduces nucleotide substitutions into a variable repeat (VR) while preserving cis- and trans- acting elements needed for future rounds of diversification. We have identified chromosomal DGRs within multiple strains of the opportunistic, human pathogen *Legionella pneumophila* (*Lp*), the etiological agent of Legionnaires' disease. *Lp* DGRs are found within horizontally acquired genomic islands that are often found within larger integrative and conjugative elements (ICE). Strikingly, *Lp* DGRs can theoretically generate ~10^{26} unique nucleotide sequences within their target genes, which upon translation could generate ~10^{19} distinct polypeptide sequences within the C-terminal C-type lectin (CLec) domains of their TPs. The *Lp* Corby TP gene, *ldtA*, encodes a surface exposed twin-arginine translocated (TAT) lipoprotein that is likely trafficked via the localization of lipoprotein (LOL) system and anchored to the outer leaflet of the bacterial outer membrane (OM) with its C-terminal variable region exposed to the extra-cellular environment. Analysis of LdtA identified a non-canonical lipobox with conserved targeting residues at +2/+3 positions. Mutagenesis of the lipobox conserved cysteine, as well as substitution of targeting residues with amino acids shown to result in sorting by Lol to the OM, all resulted in retention of LdtA in the inner membrane suggesting LdtA may be trafficked by an unusual mechanism. Related DGRs found in *Legionella* isolates have been shown to encode a small family of TPs, some with similar N-terminal

domains, which would predict their surface display using a similar mechanism. We propose all *Legionella* TPs are diversified in response to the physiological state of the host and trafficked to the surface of the bacterial cell by an unusual TAT/Lol-related mechanism. While this work represents the first characterization of a chromosomally-encoded DGR in bacteria, comparative bioinformatics predicts that lipoprotein-mediated surface display of massively variable proteins may be a common feature of many different species of bacterial DGRs

Introduction

While most bacterial retroelements are considered to contribute little to no benefit to host fitness, DGRs directly benefit their hosts by accelerating the evolution of TPs ligand binding domains [1-4]. This DGR-mediated accelerated, directed evolution of a ligand binding domain was demonstrated through analysis of the bacteriophage, BPP, which parasitizes respiratory pathogens of the *Bordetella* genus. It was observed that BPP could bind to and subsequently infect *Bordetella* cells despite their phenotypic state; a striking observation because members of this genus initiate global regulatory programs to oscillate between an environmental and an infectious phenotypic state, with each state expressing a unique set of factors upon the bacterial surface [5]. This ability to bind multiple ligands is attributed to the BPP DGR TP, *mtd*, which encodes a receptor binding protein found at the distal tips of the phage tail fibers and, through mutagenic homing, diversification of *mtd* expands the repertoire of bacterial surface ligands the phage uses for attachment [2].

The molecular mechanism by which DGRs generate nucleotide diversity in variable repeats (VR) of target genes has been well described [6, 7] and these analyses have been complemented with studies investigating the means by which generated nucleotide diversity is constrained as well as displayed within conserved domains of TPs, resulting in a balance of generated peptide diversity and scaffold stability. Structural studies of BPP Mtd and the Mtd-pertactin complex were solved and revealed that TR adenines are precisely positioned to correspond to solvent exposed residues in the ligand binding pocket of a C-terminal CLec domain, reflecting co-evolution between the genetic mechanism that generates diversity and the protein scaffold that displays it [8]. Mtd is found at the distal facet of each of BPP's six tail fibers and is positioned so that each tail fiber can have multiple interactions with its ligand. While many systems rely upon an optimal configuration between a single ligand and receptor to generate a high affinity interaction, Mtd uses multiple suboptimal tail fiber-ligand interactions to generate a multivalent avidity based interaction that results in picomolar disassociation constants [8]. Comparison of five tropic variants of Mtd revealed that their CLec domains each maintained an overall similar structural conformation and nucleotide diversification generates variable residues that are discreetly positioned in loops of this invariant backbone [9]. Furthermore, while the Mtd CLec domain maintains a relatively static nature, it is able to interact with a number of diverse epitopes, which is essential to Mtd function in binding disparate host cell ligands [8].

A second DGR TP found in the pathogenic oral spirochete *Treponema denticola*, TvpA (TDE2269), was solved and its comparison with Mtd revealed several interesting observations. While the overall structure of TvpA and Mtd appear to have little in

common, only sharing ~16% nucleotide identity and forming a homo-trimer or monomer, respectively, they both contain remarkably similar C-terminal CLec domains [10]. *T. denticola* DGR TR adenines, like the BPP element, are positioned to correspond to and diversify solvent exposed residues within the VR containing CLec domain. Superpositioning both these CLec domains revealed that many of the variable residues are structurally conserved, with additional sites of variability found in TvpA being dispersed along the rigid lectin backbone [10]. This indicates DGR TPs with greater potentials for diversity utilize a similar CLec scaffold and additional sites of variability are dispersed throughout the backbone. Comparisons of Mtd and TvpA identified a third conserved fold which has coevolved with a genetic mechanism to display massive amounts of peptide diversification, and bioinformatics suggests the CLec fold is common solution to display variability shared amongst a majority of DGR TPs [4, 8, 10].

While the *Bordetella* BPP Mtd and *T. denticola* LvpA have been paradigms for structural and functional studies of DGR TPs, recent bioinformatic studies show these elements are found in organisms with disparate lifestyles and are widely distributed in most bacterial phylum with significant sequence coverage [11, 12]. These putative DGRs encode TPs with a diverse set of domains and predicted functions. Many phage DGRs contain TPs where structural and/or genomic contextual analysis as well as homology to known proteins suggests they are phage tail proteins and likely play a similar role as Mtd (Personal communication – Mari Gingery). TPs found in the marine filamentous bacteria *Trichodesmium erythraeum* contain N-terminal domains associated with signal transduction pathways e.g. serine/threonine protein kinase or caspase-like cysteine protease domains (personal communication – Mari Gingery). *T. denticola* TvpA

does not contain domains with predicted function, but we have identified a putative signal peptide with a lipobox, suggesting it is a lipoprotein.

Here we present the first functional analysis of a bacterial chromosomal element identified in *Lp* strain Corby which encodes all components characteristically shared amongst DGRs. *Lp* is a gram-negative intracellular pathogen which has evolved strategies to evade predation in the environment, leading to accidental virulence in humans [13, 14]. We previously demonstrated the *Lp* DGR is an active element capable of directing adenine-specific mutagenesis of *ldtA*, with the potential of generating ~10^{19} distinct polypeptide sequences within the CLec domain of LdtA, and that diversification typically occurs under conditions associated with stress. Analysis of *ldtA* identified a bipartite signal peptide with an N-terminal Twin-Arginine Translocation (TAT) as well as Lipoprotein lipobox (LPP) motif, in addition to the conserved C-terminal VR-containing CLec domain. We demonstrate that this bipartite N-terminal signal peptide targets LdtA to the outer facet of the outer membrane, allowing surface display of the variable C-terminal domain to the extracellular milieu. To better understand the precise pathways required for surface display of LdtA we individually analyzed the contribution of TAT and LPP. Mutations in the TAT pathway resulted in loss of surface display. Mutagenesis of the lipobox conserved cysteine as well as replacement of targeting residues with amino acids shown to result in sorting by the localization of lipoprotein (Lol) system to the OM all resulted in retention of LdtA in the inner membrane. This suggests that LdtA is trafficked to the surface by an unusual mechanism. We had identified homologous DGRs within several clinical isolates of *Lp*. These homologous elements encode nearly identical diversification machinery, but have co-opted different carrier sequences to

display related variable domains, demonstrating the modularity of these systems. Here we present *Lp* Corby as a model system for studies of how lipoproteins can be displayed upon the bacterial surface. Furthermore, an examination of TPs in diverse gram-negative bacteria suggests that lipoprotein anchoring and surface display of DGR-diversified protein repertoires is a common theme in gram-negative bacteria.

Results

Anatomy of a *Legionella* DGR and analysis of the TP gene, *ldtA*. The DGR in Figure 1 is located within a ten kilobase pair (kbp) genomic island in the chromosome of *Lp* strain Corby. Flanking genes are predicted to encode a heavy metal transport system (*helA-C*) and a type IV secretion system with associated regulators (*lvrA-C*) [15]. The island shows signs of recent horizontal acquisition, as indicated by a G+C content (45%) that differs from the rest of the genome (38%), and the presence of a transposase and an unrelated RT [16]. The retroelement itself encodes a DGR-type RT [3], an Avd homolog, cognate TR and VR sequences that differ at sites corresponding to adenines in TR, and tandem stem-loop/cruciform structures downstream of VR [6, 7]. The 148 bp TR contains 43 adenines which most often occupy the first two positions of AAC or AAT codons, allowing maximal amino acid diversity while excluding the possibility of nonsense mutations generated by adenine-mutagenesis. Following mutagenic homing, the *Lp* TR can theoretically generate 4^{43} (~10^{26}) unique DNA sequences capable of encoding ~10^{19} different polypeptides, a repertoire of massive

proportions. Diversification of *ldtA* has been observed under *in vitro* conditions using an overexpression system, as well as *in vivo* conditions.

Analysis of the Corby DGR TP gene identified similarities to as well as distinctions from known DGR target genes. VR encoded sequences are located at the C-terminus of LdtA, within a domain predicted to adopt a CLec fold similar to that of Mtd (Sup. Figure 1). While no other conserved domains were identified, *in silico* analysis followed by manual curation identified a putative bipartite signal peptide containing two protein trafficking domains at the N-terminus of LdtA [17]. The first predicted domain is a TAT motif of amino acids KKRHFFR that differs from the consensus in *E. coli* of (S/T)RRXFLK, where X is any amino acid, but is characteristic of known and putative TAT substrates in *Lp* [18, 19]. The second is a lipobox motif of FFSC that is also non-canonical compared to the standard motif of LVI-ASTVI-AGS-C [20, 21]. The TAT pathway is an alternative secretion system found in plants and bacteria that can translocate folded proteins or protein complexes across lipid bilayers [22] and lipobox motifs mediate signal peptide cleavage, lipid modification, and anchoring to the inner or outer membrane [23]. Although the ability of TAT and lipobox secretion motifs to function in concert has not been thoroughly characterized [24, 25], we hypothesized that the N-terminus of LdtA mediates secretion, membrane localization, and potentially surface exposure.

LdtA is surface exposed. *Lp* Corby cells expressing LdtA or control proteins with C-terminal hemagglutinin (HA) epitope tags were lysed and separated into soluble and membrane fractions. LdtA partitioned to the membrane fraction along with the

known outer membrane protein macrophage infectivity potentiator (MIP) [26] and DotA (IM protein control) [27] while RecA appeared in the soluble fraction (Figure 2A) [28]. On further separation using isopycnic sucrose gradient ultracentrifugation, DotA was enriched in fractions containing IM proteins while MIP and LdtA preferentially partitioned to the fraction containing OM proteins (Figure 2B). To determine the orientation of LdtA, intact bacterial cells were treated with proteases under conditions that preferentially degrade surface-exposed proteins (Figure 2C). Cells were induced to express HA-tagged LdtA, MIP, DotA, RecA, or IcmX as a periplasmic control [29] and incubated with increasing concentrations of proteinase K [30]. MIP, an integral OM protein, showed moderate protease sensitivity while periplasmic, IM, and cytoplasmic control proteins were relatively unaffected. In contrast, LdtA was highly sensitive to protease treatment, as indicative of surface localization.

Indirect immunofluorescence was used as an independent approach to test surface exposure. Cells expressing full length LdtA-HA were recalcitrant to visualization attempts, and we hypothesized this was due to sequestration of the epitope tag within the CLec folded structure. Based on structural modeling predictions we constructed a variant that expressed the first 370 amino acids of LdtA fused to a triple-HA C-terminal epitope tag (LdtA-370-3HA). Surface immunofluorescence was readily detected using intact cells, while visualization of the IcmX or DotA negative controls required OM permeabilization (Figure 2E). In summary, membrane fractionation, protease sensitivity, and surface immunofluorescence support the conclusion that LdtA is an OM protein with a surface-exposed C-terminus.

LdtA is a TAT-secreted lipoprotein. Translocation through the TAT pathway requires cytoplasmic recognition by the TatABC complex of signal sequences with the consensus motif SRRxFLK, which is conserved in plants and other bacteria but more variable in *Legionella* [18, 19, 22]. The predicted LdtA TAT secretion motif retains requisite recognition components including polar N-terminal residues, a hydrophobic core with the arginine that is absolutely required for TAT transport, and less hydrophobic C-terminal residues (Figure. 1). To determine if the TAT system is required for LdtA translocation across the IM, we generated an in-frame deletion in *Lp* Corby *tatB* [31, 32]. LdtA-HA was protease resistant in the Δ*tatB* strain and protease sensitivity was restored by complementation with *tatB* (Figure 2D). Confirmatory results were obtained by immunofluorescence (Figure.23E).

Having shown that LdtA is secreted through the TAT pathway, we were curious to determine how it becomes localized in the OM. TAT motifs often contain a consensus sequence at their C-terminus, Ala-Xaa-Ala, which mediates cleavage by signal peptidase I [18]. In contrast, we identified a potential lipobox motif, FFSC, at the analogous position in LdtA (Figure 1A). This suggested a hybrid signal peptide that combines TAT translocation with lipoprotein signal peptide cleavage, lipid modification of the conserved cysteine (Cys-20), and trafficking to the OM by the LOL system. To test this hypothesis, we first determined if the lipobox conserved cysteine (Cys-20) is required for surface exposure. As shown in Figure 2D, alanine (C20A) or serine (C20S) substitutions at this site rendered LdtA-HA protease resistant, with a corresponding loss of surface immunofluorescence (Figure 2F). To determine the step at which trafficking is blocked, we compared the effects of OM permeabilization on phenotypes of Cys-20

substitutions in *wt* vs. Δ*tatB* mutants. Wild type *Lp* expressing LdtA-C20A showed sensitivity to proteinase K under permeabilizing conditions but not in intact cells, as observed with the periplasmic control protein IcmX (Figure 2C-F). In contrast, LdtA-C20 was protease resistant in the Δ*tatB* strain, even after OM permeabilization. Immunofluorescence results paralleled protease sensitivity assays (Figure 2E, F) and showed that mutation of Cys-20 causes missorting to the periplasm, a phenotype that would be predicted to result from the lack of modification of an OM lipoprotein.

Bacteria often modify lipobox cysteines with palmitic acid by linkage to *N*-acyl-*S*-diacylglyceryl-Cys moieties [20, 21]. Acylbiotin-exchange chemistry, a sensitive alternative to labeling cells with radioactive fatty acids, was used to detect post-translational modification of LdtA (Sup. Figure 2A, B) [33]. Protein lysates from *E. coli* and wild type or Δ*tatB Lp* cells expressing LdtA or LdtA-C20S were treated with hydroxylamine to remove acyl-linked lipids which were then replaced with biotin. Biotinylated proteins were column purified, eluted, and analyzed by western blotting. Biotin labeled LdtA was found in wild type *Lp* but not in Δ*tatB* mutants and LdtA-C20S was refractory to biotinylation (Sup. Figure 2B). Taken together, our results show that after translocation across the IM by TAT, LdtA is anchored in the outer surface of the OM by an acyl-linked lipid modification.

The LdtA signal peptide is sufficient for surface localization. We wanted to assess if the LdtA signal peptide was sufficient to display a protein on the surface of a gram-negative bacterium. The first twenty-six amino acids of *ldtA* were fused in-frame with a signal peptide-less green fluorescent protein (gfp) and was expressed from a

plasmid vector under the control of the *tac* promoter (pldtA-gfp) [34]. Wild type *Lp* cells expressing pldtA-gfp were treated with exogenous proteases, as above, and showed similar sensitivity as full length epitope tagged LdtA (Figure 3A). In contrast, *Lp* cells expressing full length gfp (data not shown) or *LpΔtatB* cells expressing pldtA-gfp showed no sensitivity to exogenous proteases which is consistent with previous data and demonstrates the LdtA bipartite signal peptide is sufficient to traffic a carrier peptide for surface display in *Lp*. Immunofluorescence gave complementary results (Figure 3B).

Contribution of lipobox targeting residues to surface localization of LdtA.

Lipoprotein processing is dependent upon the recognition of lipobox amino acids by cellular machinery found at the IM/periplasmic interface. The four amino acid lipobox is recognized by Lgt and the conserved cysteine is modified with a single diacylglycerol moiety. The modified signal peptide is then cleaved by the signal peptidase II (LspA), resulting in the conserved cysteine being the first reside or +1. This cysteine is then further processed with the addition of an N-acyl moiety by Lnt [20] and this final modification with a lipid moiety is thought to ensure retention within lipid bilayers. Processed lipoproteins are then either maintained in the IM or trafficked to the OM based on interaction between the Lol system and the cleaved/modified signal peptide. Recent experiments have demonstrated that trafficking by LOL depends upon the chemical characteristics of the conserved cysteine as well as the following 2-3 residues, commonly referred to as the +2 residue or +3 residue, respectively, and will be referred to as targeting residues [20, 35].

We have demonstrated the requirement of the TAT translocon for secretion of LdtA across the *Lp* IM as well as post-translational modification of the lipobox's conserved cysteine which is consistent with LdtA being processed by *Legionella* homologues of *lgt* and *lnt*. As the Lol system is the only known system capable of trafficking lipoproteins from the bacterial IM to the OM, we sought to investigate its contribution to surface display of LdtA. Since the Lol system and the contribution of targeting residues have been best characterized in *Escherichia coli* we chose a *wt* strain, MG1655, for our analysis [20, 36-38]. The plasmid vector pldtA-gfp or derivatives with amino acid replacement of the conserved cysteine (+1 residue) with alanine, replacement of the +2 residue with aspartate (D) or serine (S), or replacement from the +2 to +4 residues with leucine-methionine-leucine (LML) were transformed into MG1655, induced for expression and treated with exogenous proteases, as above, under conditions which preferentially degrade surface-exposed proteins. Cells expressing the *wt* signal peptide showed sensitivity to proteases that was consistent with the expression of full length LdtA in *Lp* (Figure 4A) however amino acid substitution of targeting residues resulted in loss of sensitivity to exogenous proteases, indicative of loss of surface display. Immunofluorescence results paralleled protease sensitivity assays (Figure 4B) and showed that while gfp could be detected from all cells, only cells expressing LdtA-egfp could be counter-stained with anti-GFP antibodies.

Having observed loss of surface exposure, we wanted to determine to which membrane the mutants were being trafficked. *E. coli* cells expressing pLdtA-egfp or amino acid replacement derivatives were lysed and separated into soluble and membrane fractions, as above. The total membrane fraction was further separated into

inner- or outer-membrane fractions using isopycnic sucrose gradient ultracentrifugation. LdtA-gfp, as well as the outer membrane control proteins Lpp and DotA, was enriched in sucrose fractions corresponding to the OM. Interestingly, every amino acid replacement derivative tested was enriched in the fractions corresponding to IM proteins (Figure 4C), and analysis using antibodies specific of outer membrane porin (omp) F revealed no alteration in protein trafficking (Figure 4D), demonstrating altered trafficking was specific. These replacement residues were chosen because they had been demonstrated to be crucial for targeting of lipoproteins to specific membranes. Substitution of the +1 residue with aspartate is classically considered as the Lol avoidance signal and results in IM retention, while in contrast many lipoproteins with serine at the +1 position are found in the OM [20].

Cleavage of the LdtA signal peptide is dependent upon the +2 residue. We have demonstrated that substitutions of lipoprotein targeting residues in LdtA with residues previously demonstrated to result in trafficking of the substrate to the OM resulted in retention of LdtA in the IM. This retention could be from lack of recognition by LolCDE or that targeting residues which mediate surface display of LdtA cannot be altered. To investigate these possibilities we assessed one of the initial steps in lipoprotein maturation, the cleavage of the signal peptide by signal peptidase II, LspA. Cells expressing LdtA and the targeting residue replacement derivatives were induced for expression and whole cell pellets were probed by western blot using commercially available antibodies. When comparing LdtA-gfp which is displayed upon the bacterial surface, with C20A which is retained in the IM, we observed that they migrated as the same size protein. This is interesting as the conserved cysteine (C20) is necessary for

recognition of a lipoprotein by Lnt and LspA (Figure 5). Furthermore, replacement of the +2 residue with serine displayed similar migration as *wt* and C20A LdtA but L21D as well as the triple replacement LML::AGT appeared to migrate at a height consistent with an uncleaved signal peptide. These observations are inconsistent with cleavage of the LdtA signal peptide by LspA.

Distribution of TPs in *Lp* isolates. We had previously identified several putative DGRs within a library of clinical isolates as well as several elements within deposited nucleotide sequence databases. The *Lp* Corby encodes LdtA with a bipartite N-terminal signal peptide containing non-canonical trafficking motifs and a conserved, VR containing C-terminal CLec domain. The *Lp* D5572 DGR encodes a predicted TP (LdtB) which shares ~93% homology with LdtA on the amino acid level as well as the same bipartite signal peptide. A third DGR in *Lp* strain D5591 encodes a TP (LdtC) which shares ~40% homology with LdtA on the amino acid level, but shares ~88% homology to a TP (LdtD) found in a fourth DGR within *Legionella tunisiensis* (Figure 6). LdtC and LdtD share little N-terminal homology to LdtA and do not contain the same signal peptide however, *in silico* analysis predicts they might be secreted to the cell surface via a different pathway [39-41]. *Legionella* TPs are of similar size, ranging from 502 to 512 amino acids. While they are phylogenetically bifurcated, *Legionella* TPs show a high degree of amino acid homology at residues corresponding to the beginning and end of the VR which are likely necessary for stability of the conserved CLec domain. While these scaffold stabilizing residues are conserved, each DGR target gene contains a VR with patterns of adenine mutagenesis that, upon translation, encode for unique peptide sequences. This demonstrates that the same scaffold supports different

polypeptide patterns and diversification is likely a response to a stimulus which has been selected for.

LdtA as a model for diversified surface lipoproteins. The ability to introduce massive amounts of diversity in surface-anchored bacterial proteins seems likely to be an adaptive trait. In support of this idea, a partial survey of sequenced bacterial genomes revealed multiple DGRs that are predicted to diversify TPs with N-terminal lipoprotein sequences. As shown in Figure 7, predicted DGR-diversified TPs in *T. denticola, Bacteroides fragilis*, *B. thetaiotaomicron*, *Vibrio augustum,* and *Shewanella baltica* contain N-terminal LPP motifs. Although LdtA is known and LdtB is predicted to be TAT-secreted, the other TPs in Figure 6 have SEC secretion signals. *T. denticola* TPs are particularly interesting. We predict they are diversified by a single DGR and the majority includes highly predicted LPP processing signals. Spirochetes are known to make extensive use of the LPP secretion pathway to anchor proteins to their outer surface [30]. Taken together, our results suggest that lipid modification and surface exposure are conserved features of proteins that are diversified by bacterial chromosomal DGRs.

Discussion

We have demonstrated that LdtA traffics across the inner membrane via the TAT translocon, is cleaved, lipidated, and localized to the external face of the outer membrane, presumably by the LOL system. Although the manner in which TAT translocation and LOL processing systems intersect, and the mechanism through which

lipoproteins are resolved to the outer leaflets of outer membranes remain as open questions [20, 25], our results clearly demonstrate that LdtA adorns the surface of *Lp*. TAT secreted lipoproteins are a poorly understood class of proteins which are broadly found in a diverse number of bacteria. For certain organisms TAT-lipoproteins are predicted to comprise a majority of encoded proteins, and they have been suggested as facilitating host niche adaptation [42]. We show that lipid modification of LdtA as well as chemical characteristic of the lipoprotein targeting residues correlates with surface exposure. Cleavage of LdtA signal peptide appears to be dependent upon the +2 residue and, although further analysis is required, it is an interesting observation as previous reports identified the lipobox -1 residue as having an effect on substrate cleavage by signal peptidase II [43]. The molecular means explaining why replacing an aliphatic residue (leucine) with a structurally and chemically similar residue (alanine) has such a dramatic effect on protein localization needs further investigation. It is unclear if these observations are unique for LdtA, DGR TPs, or for TAT secreted lipoproteins as a class. Furthermore, we expressed and detected LdtA upon the surface of both *Lp* as well as *E. coli*, suggesting the ability to traffic and display TAT-lipoproteins upon the bacterial outer membrane involves pathways conserved among gram negative proteobacteria.

We demonstrated that LdtA is anchored to *Lp* outer membrane, likely through its N-terminal domain, and that the C-terminal CLec domain is available to the extracellular milieu. CLec domains have been reported as general ligand binding domains found in metazoan, bacterial, and viral proteins, and known binding partners include other proteins, sugars, lipids, and inorganic ligands [44]. While the function of LdtA is

presently unknown, a likely possibility is that it facilitates *Lp* receptor-ligand interactions of importance for survival in the environment and/or interactions with host cells.

The *Lp* Corby, D5591, likely the D5572 and the *Lt* LegM DGRs are located within conserved genomic islands that bear the hallmarks of recent horizontal acquisition. Our analysis of DGRs in clinical and environmental isolates revealed that all *Legionella* DGRs share nearly identical diversification machinery: *avd*, TR, and *RT* loci as well as highly conserved recognition elements (IMH and stem loops). Despite this similarity, each VR displays a strain specific pattern of adenine mutagenesis which would translate into a unique polypeptide sequence within the conserved CLec domain. This is consistent with the hypothesis that these elements are active in nature and that target protein functions are under selection. Additionally, it has been reported that the DGR containing ICE element is capable of being transferred intra- as well as inter-species, providing a means for dissemination of DGRs throughout bacterial populations [45]. The observation that similar VRs have been fused to entirely different N-terminal sequences in *Lp* D5591/*Lt* LegM vs*. Lp* Corby/D5572 and that all TPs are predicted to be localized on the bacterial surface provides further illustration of the modular nature of diversified proteins and the versatility of the VR containing CLec scaffold. The widespread distribution of DGRs in nature, both within *Legionella* species as well as within a wide range of bacteria, and their adaptation to mediate both phage and bacterial surface display is not surprising given their utility as internally programmed, self-renewable systems capable of accelerating the evolution of adaptive traits.

Materials and Methods

Bacterial Strains, Growth, and Mutant Construction. *Lp* Corby [14] was a kind gift from Natalia Kozak from the Centers for Disease Control and Prevention (CDC). *Lp* Corby and derivatives were routinely maintained in culture in yeast extract (PYG) broth or on buffered charcoal-yeast extract (BCYE) media as previously described [15]. In-frame deletions and substitution mutations were constructed using allelic exchange with the *sacB* negative selection marker on BCYE agar containing 7.5% sucrose [32]. *Lp* Corby gene loci targeted for mutational analysis were grown on BYCE + Kanamycin (Km) 10μg/mL included *tatB* (LpC_3208), *ldtA* (LpC_1853), *avd* (LpC_1854), *RT* (LpC_1855) and intergenic regions between *ldtA* and *avd* or *avd* and *RT* representing stem/loops, VR, or TR, respectively. The broad host vector pMMB208 was used for complementation of mutants and protein overexpression and cells harboring this vector were grown in media supplemented with 5μg/mL chloramphenicol (Cm) [34].

Plasmid Construction. For protein expression studies, *MIP*, *dotA*, *icmX*, *recA*, *ldtA*, or were fused at their 3' ends to sequences encoding HA epitope tags, cloned into pMMB208 under the control of the *tac* promoter and induced for protein expression by the addition of Isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM to the growth media. For studies involving the LdtA signal peptide sequence fused to egfp, primers were used to amplify from the start codon to the twenty-sixth amino acid of LpC_1853 and this product was used to amplify from the twenty-sixth amino acid of egfp (Clonetech) to its stop codon. This fusion was cloned into pMMB208 and induced for protein expression as above. *LdtA* fusions with mutated targeting

residues were generated using PCR with primers harboring the desired mutations, cloned into pMMB208 and induced for protein expression, as above.

PCR Screen for DGR Genes. To screen *Lp* strains for putative DGR genes, a series of degenerative primers were designed using previously described methods (36). The first iteration of primers used is as follows: 16SFor- AGCATKGTCTAGCTTGCTAG, 16SRev- TCCTCCCCACTGAAAGTG, avdFor- TGTTTGAGGTAACGAAAGATTTC, avdRev- CCGGTCACCTGCTTGCCTA, RTFor- AAATCATCGACGTAACGACCATA, RTRev- CTTTCGTGACCGTGTGGTGC.

Sequencing of *Lp* Strains. The genomes of *Lp* strains D5549, D5572, and D5591 were assembled and sequenced on an Illumina platform as previously described [46]. Reads were assembled into contigs using Assembler and viewed with Artemis.

***In silico* analysis of deposited nucleotide sequences.** The National Center for Biotechnology Information (NCBI) nucleotide sequences, both nucleotide collection and whole-genome shotgun contigs, were analyzed with the build in blast program using the *Lp* Corby DGR TR as the query or with the blast program using the Corby DGR RT amino acid sequence as the query. Results were constrained by limiting queries to the *Legionella* genus. Whole-genome shotgun contigs with putative DGR elements were downloaded, viewed, and manually curated using Vector NTI, commercially available software (Invitrogen). Analysis of DGR target protein for conserved protein domains or conserved protein secretion domains was performed using TatP 1.0 [41], LipoP 1.0 [40], PSORTb [39], and SignalP4.1 [47].

Protease Sensitivity Assays and Immunofluorescence Microscopy. *Lp* cells expressing genes of interest for localization studies were sub-cultured in PYG to an OD_{590} of 0.2, grown for four hours, and induced for protein expression by the addition of IPTG to a final concentration of 1 mM for four hours. Cells were harvested by centrifugation and washed with phosphate buffered saline (PBS) pH 7.3 supplemented with 5mM MgCl. Cells were normalized by OD and treated with 0, 50, 100, or 200 mg/mL of proteinase K (Sigma) for one hour at room temperature. Cells were harvested by centrifugation and washed with PBS+5mM MgCl. Cell pellets were solubilized, proteins were separated by SDS-PAGE and proteins of interest were detected by western blot using anti-Myc (Covance), anti-HA (Abcam), or anti-*Lp* (Abcam) antibodies.

For whole-cell immunofluorescence microscopy, aliquots of untreated or proteinase K treated cells washed with PBS + 5mM MgCl were dropped onto gelatinized glass slides and incubated at room temperature for 30 minutes. Unbound cells were removed and adherent cells were fixed with 100 µL PBS + 4% paraformaldehyde for 30 minutes at room temperature. Fixed cells were washed twice with PBS and then blocked by addition of 100 µL PBS + 5% BSA for one hour at room temperature. Once blocked, cells were incubated with commercially available anti-Myc, anti-HA, anti-GFP or anti-*Lp* (1:250) antisera in 100 µL PBS + 5% BSA for one hour at room temperature. Cells were washed three times in PBS and then incubated in 100 µL PBS + 5% BSA with species-appropriate fluorophore conjugated secondary antibody for one hour at room temperature. Cells were washed three times with PBS, coverslips mounted by addition of vectashield (Vector Laboratories), and coverslips were sealed with nail

polish. Microscopy was performed using an AX10 Imager Z-10 (Carl Zeiss) and analysis performed using Axiovision.

Isolation of Membrane Fractions and Acylbiotin-Exchange Chemistry. Isolation of *Lp* Corby membrane proteins was performed as previously described [30] using 100 mL cultures and a discontinuous sucrose cushion from 30% to 60% sucrose density. Membrane fractions of *Lp* Corby cells expressing *wt* or mutant LdtA were isolated as above and post-translation modification of the lipobox cysteine determined through acyl biotin-exchange chemistry as previously described with minor modifications [33]. Briefly, membrane proteins were suspended in PBS + protease inhibitor/5mM EDTA/1% Triton X-100 buffer (buffer 1) which was supplemented with N-ethylamaleimide to a final concentration of 50 mM for 30 minutes at 4°C. Proteins were precipitated with methanol/chloroform on ice for five minutes. Precipitated proteins were washed and suspended in 100 µL of PBS + 5mM EDTA supplemented with hydroxylamine, when indicated, and biotin-BMCC (Pierce) then incubated at 4°C for one hour. Proteins were precipitated with methanol/chloroform on ice for five minutes and then suspended in buffer 1. Biotin labeled proteins were purified by column affinity purification using streptavidin beads, washed, eluted by boiling in protein running buffer, and analyzed by western blotting.

Protein Structure Prediction. LdtA, LdtB, and LdtC were subjected to three-dimensional structure prediction using the algorithm in Phyre 2 [48]. In all cases, no predictions were made for the N-terminal regions of these proteins but predictions with high confidence were made for the C-terminal ~150 residues. These predictions

indicated that the C-terminal domains of LdtA, LdtB, and LdtC have CLec-folds. The highest scoring templates in each case were a *Bacteroides ovatus* hypothetical protein (BACOVA_04982, PDB 4EPS) and *E. coli* intimin (PDB 1E5U) [49]. The latter has been characterized as CLec-fold proteins; inspection of the former indicates that it has a CLec-fold as well. For LdtA, the confidence level to BACOVA_04982 and intimin was 99.9% and 94.0%, respectively, for LdtB it was 99.7% and 92.7%, respectively, and for LdtC it was 99.5% and 86.6%, respectively. The template for modeling the structures of the C-terminal domains of LdtA, LdtB, and LdtC was BACOVA_04982 in each case. The LdtA model has a 99% confidence level (except at eight residues), the LdtB model a 99% confidence level (except at eight residues), and the LdtC model a 99% confidence level (except at ten residues).

Figure Legends

Figure 1. . A *Legionella pneumophila* DGR. A DGR in *Lp* strain Corby is found within a ten kbp genomic island (green dashed box) adjacent to a heavy metal transport gene cluster (*helA-C*) and Type IV secretion system regulatory genes (*lvrR-C*, purple). The DGR-containing genomic island has a higher G+C content (percentile score) than flanking regions, a non-DGR RT (dark blue), genes annotated as hypothetical proteins (grey), and putative transposase (orange). DGR loci are expanded with the TP gene (*ldtA*), predicted secretion (TAT) and localization sequences (Lpp), accessory protein gene (*avd*), and DGR-encoded reverse transcriptase (*RT*) identified. Additional DGR elements: DNA stem/loops, IMH, and IMH* are indicated. The LdtA signal peptide (green box) contains a polar N-region (residues 1-11), followed by a hydrophobic core (residues 12-15), and a non-polar C-region (residues 17-23). A non-canonical TAT motif, KKRHFFR, is predicted which differs from the *E. coli* consensus of (S/T)RRXFLK but retains requisite arginine (R6). The LdtA signal peptide contains a lipobox motif, FFSC, which is predicted to be lipid-modified. A C-terminal CLec domain with homology to *Bb* BPP Mtd is identified.

Figure 2. LdtA is a surface exposed lipoprotein. Epitope (HA) tagged LdtA or control proteins were used for sub-cellular localization experiments. (*A*) Cells were lysed by French press and cellular constituents separated by high speed centrifugation. Proteins with known sub-cellular locations: MIP (OM, 20), DotA (IM, 21), and RecA (cytoplasmic) were compared with LdtA. RecA was enriched in the soluble fraction while LdtA, MIP

and DotA were enriched in membrane fractions. Membrane proteins detected in soluble fractions are due to overexpression and increased pools of secretion intermediates. (*B*) High speed isopycnic sucrose gradient fractionation was used to separate membrane proteins and fractions were probed by western blotting with anti-HA antibodies. Fractions corresponding to IM or OM are shown. (*C*) *Lp* Corby cells expressing epitope tagged proteins were treated with increasing concentrations of proteinase K (0-200µg/mL) to digest surface-exposed proteins and whole-cell lysates were probed as above. HA-tagged IcmX, a periplasmic T4SS component, was included as an additional control. (*D*) Top: Mutation of the essential Cys in the LdtA Lpp processing signal to Ala (C20A) or Ser (C20S) protected LdtA from digestion by proteinase K. Middle: An in-frame deletion in *tatB* (Corby∆*tatB*) eliminated LdtA digestion, and complementation by plasmid-expressed *tatB* (p*tatB*) restored protease sensitivity. Bottom: *Lp* Corby cells expressing epitope tagged IcmX, LdtA C20A, or Corby∆*tatB* expressing LdtA were permeabilized with EDTA/Lysozyme and treated with increasing concentrations of proteinase K. (*E*) Indirect immunofluorescence detection of epitope-tagged surface proteins in *Lp* Corby. Cells were treated with anti-HA or anti-*Lp* antibodies under permeabilizing (EDTA and lysozyme) or non-permeabilizing conditions, as indicated. (*F*) Top: Indirect immunofluorescence detection of epitope-tagged LdtA or LdtA C20S mutant derivative under permeabilizing or non-permeabilizing conditions. Bottom: Indirect immunofluorescence detection of epitope-tagged LdtA in *Lp* Corby ∆*tatB* in the presence or absence of a complementing plasmid (p*tatB*).

Figure 3. LdtA signal peptide is sufficient for protein surface display in *Lp*. (*A*) The first 26 amino acids of *ldtA* was fused in-frame with *gfp* and expressed in *Lp* Corby cells. Corby and Corby*Δtat*B cells expressing fusion proteins were treated with increasing concentrations of proteinase K (0-200μg/mL) to digest surface-exposed proteins and whole-cell lysates were probed as above. Protein ladder (L) is indicated. (*B*) Indirect immunofluorescence detection of fusion proteins in *Lp* Corby strains. Corby and Corby*Δtat*B cells expressing fusion proteins were treated with anti-GFP antibodies (red) under non-permeabilizing conditions, as above. Detection of gfp fluorescence using a standard 488 filter.

Figure 4. The contribution of lipobox targeting residues to LdtA-derived surface display of proteins in *E. coli*. (*A*) *E. coli* cells expressing pldtA-gfp and targeting residue substitution derivatives, as indicated, fusion proteins were treated with increasing concentrations of proteinase K (0-200μg/mL) to digest surface-exposed proteins and whole-cell lysates were probed using anti-GFP antibodies, as above. Expression of full length enhanced GFP (Clonetech) was used as an additional control. (*B*) Indirect immunofluorescence detection of fusion proteins in *E. coli*. Cells expressing pldtA-gfp and targeting residue substitution derivatives were treated with anti-GFP antibodies (red) under non-permeabilizing conditions, as above. (*C*) High speed isopycnic sucrose gradient fractionation of LdtA-GFP and targeting residue substitution derivatives was used to separate membrane proteins and fractions were probed by western blotting with anti-GFP antibodies. Fraction numbers with corresponding sucrose

density are shown. Fractions corresponding to IM or OM are underlined. (*D*) High speed isopycnic sucrose gradient fractionation of strains in *C* was used to separate membrane proteins and fractions were probed by western blotting with anti-OmpF antibodies as a control for trafficking of proteins.

Figure 5. Cleavage of the LdtA signal peptide depends on the characteristic of +2 residues. Whole cells lysates of *E. coli* cells expressing pldtA-gfp and targeting residue substitution derivatives were analyzed with anti-GFP antibodies for cleavage of signal peptide. Cells expressing empty vector (EV), pldtA-gfp (ldtA), or pldtA-gfp with amino acid substitutions in the signal peptide are indicated.

Figure 6. *Legionella* genus encodes for a family of DGR TPs. Graphical representation of the four identified DGR TPs identified in *Lp* and *Lt*. Signal peptide amino acid sequences are identified, as are conserved motifs. All strains contain conserved CLec domains, VR, and IMH. While LdtA and likely LdtB are TAT-lipoproteins, LdtC and D are predicted to be secreted by an alternative mechanism.

Figure 7. Putative DGR target proteins in diverse bacteria are predicted lipoproteins. Species and protein ORFs are indicated. Analysis of putative TPs identified N-terminal lipoboxes (red) and C-terminal VR sequences (colored arrows) found within predicted CLec domains (white box). *T. denticola* TPs are predicted to be

diversified *in trans* by a single DGR encoding TDE2269, TR, *hrdc* (an *avd* homolog), and *RT* genes. All other TPs are encoded within their cognate DGRs. Arrows represent potential nucleotide diversity generated by mutagenic homing during transfer of sequence information.

Sup. Figure 1. Predicted structures of the C-terminal domains of LdtA, LdtB, and LdtC in ribbon representation. α-helices (red), β-strands (blue), loops (gray), and the locations of VR residues are indicated. The core secondary structure elements (the paired β1β5 strands, the connecting α1 and α2 helices, and the β2β3β4 sheet) of the CLec-fold are labeled. Other secondary structures may form the inserts often found in CLec-folds.

Sup. Figure 2. . LdtA trafficking and modification. (*A*) Post-translational modification of the conserved lipobox cysteine (C20) in LdtA. *E. coli* cells expressing epitope tagged LdtA or LdtA-C20S were induced for protein expression. Cells were disrupted and total protein lysate was blocked with *N*-ethylamine followed by cleavage of post-translational modifications of cysteine by hydroxylamine, as indicated. Modifications removed by cleavage were replaced with biotin-BMCC (Pierce) and protein lysates run over a streptavidin column to bind biotinylated proteins, eluted, run on a protein gel, and probed by western blotting with anti-Myc antibodies. Differential labeling of wt LdtA and the cysteine mutant C20S identifies post-translational palmitoylation of the lipobox conserved cysteine (red arrow ~54kDa). A non-specific

product is indicated by a black arrow. (*B*) *Lp* Corby cells expressing epitope tagged LdtA, LdtA-C20S, or Corby*ΔtatB* cells expressing LdtA were induced for protein expression and total membrane proteins isolated as in (A). Unmodified cysteines in membrane proteins were blocked with *N*-ethylamine followed by cleavage of post-translational modifications by hydroxylamine, as indicated, and replacement of the modification with biotin-BMCC (Peirce). Labeled proteins were run over a streptavidin column, washed, eluted, and run on a protein gel for analysis by western blot. Lysates were probed with anti-Myc antibodies to detect epitope tagged LdtA (green) or fluorophore conjugated streptavidin detected biotinylated proteins (red). Corby*ΔtatB* was complemented by expression of *tatB* from plasmid. Red bands observed in membrane proteins without hydroxylamine (left) treatment are biotinylated proteins which have non-specifically bound and been eluted from the streptavidin column.

Figure 1.

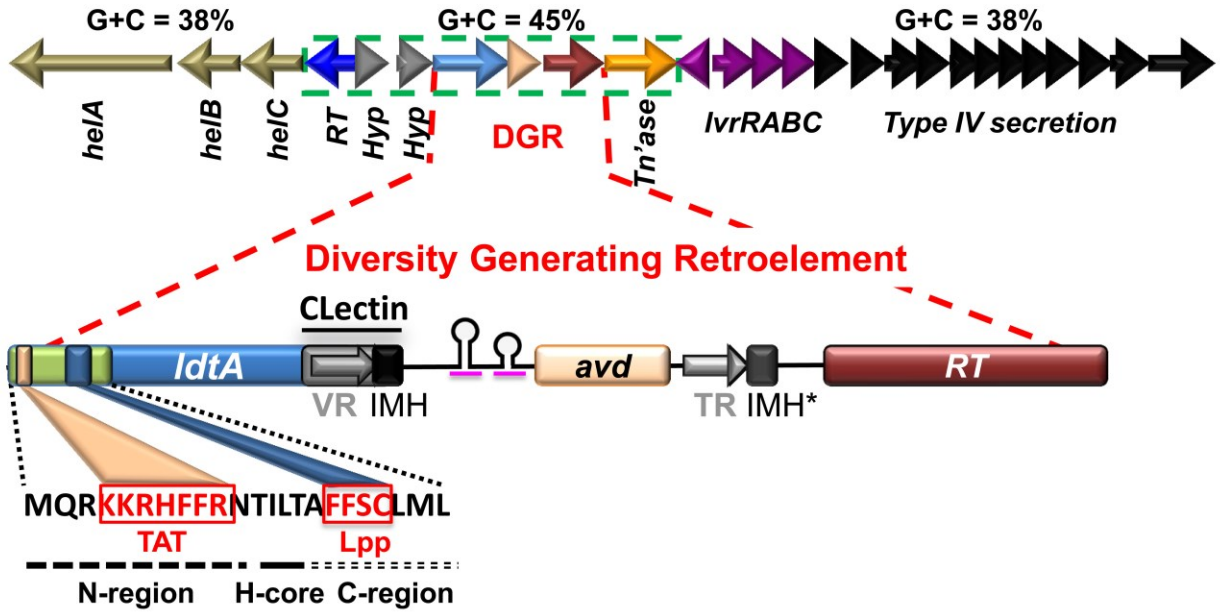

G+C = 38%
G+C = 45%
G+C = 38%
helA
helB
helC
RT
Hyp
Hyp
DGR
Tn'ase
IvrRABC
Type IV secretion
Diversity Generating Retroelement
CLectin
IdtA
VR
IMH
avd
TR
IMH*
RT
MQRKKRHFFRNTILTAFFSCLML
TAT
Lpp
N-region
H-core
C-region

Figure 2.

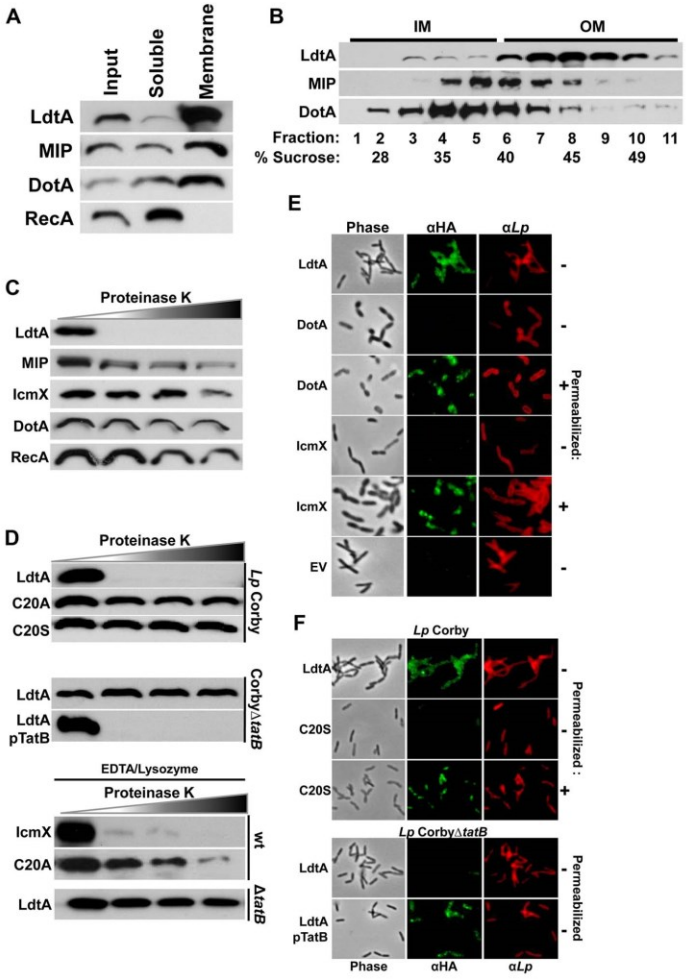

A
Input
Soluble
Membrane
LdtA
MIP
DotA
RecA
B
IM
OM
LdtA
MIP
DotA
Fraction: 1 2 3 4 5 6 7 8 9 10 11
% Sucrose: 28 35 40 45 49
C
Proteinase K
LdtA
MIP
IcmX
DotA
RecA
D
Proteinase K
LdtA
C20A
C20S
Lp Corby
LdtA
LdtA pTatB
Corby ΔtatB
EDTA/Lysozyme
Proteinase K
IcmX
C20A
wt
LdtA
ΔtatB
E
Phase
αHA
αLp
LdtA
DotA
DotA
IcmX
IcmX
EV
Permeabilized:
F
Lp Corby
LdtA
C20S
C20S
Permeabilized:
Lp CorbyΔtatB
LdtA
LdtA pTatB
Permeabilized
Phase
αHA
αLp

Figure 3.

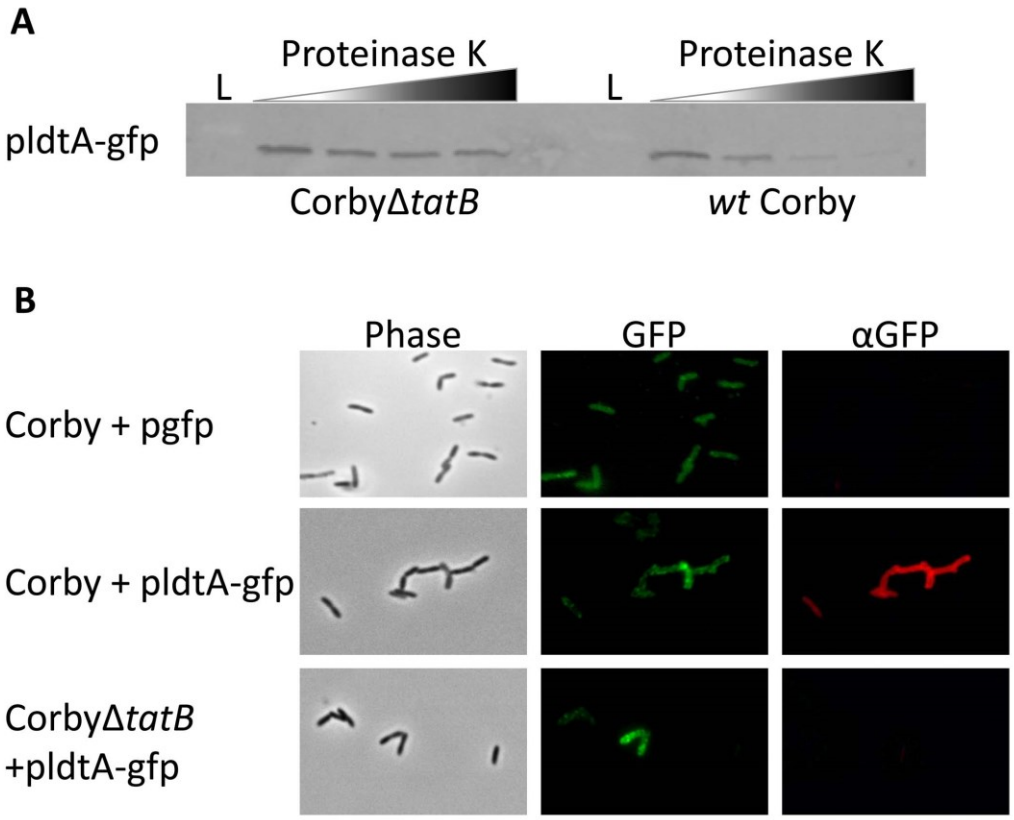

A
Proteinase K
L
Proteinase K
L
pldtA-gfp
CorbyΔtatB
wt Corby
B
Phase
GFP
αGFP
Corby + pgfp
Corby + pldtA-gfp
CorbyΔtatB
+pldtA-gfp

Figure 4.

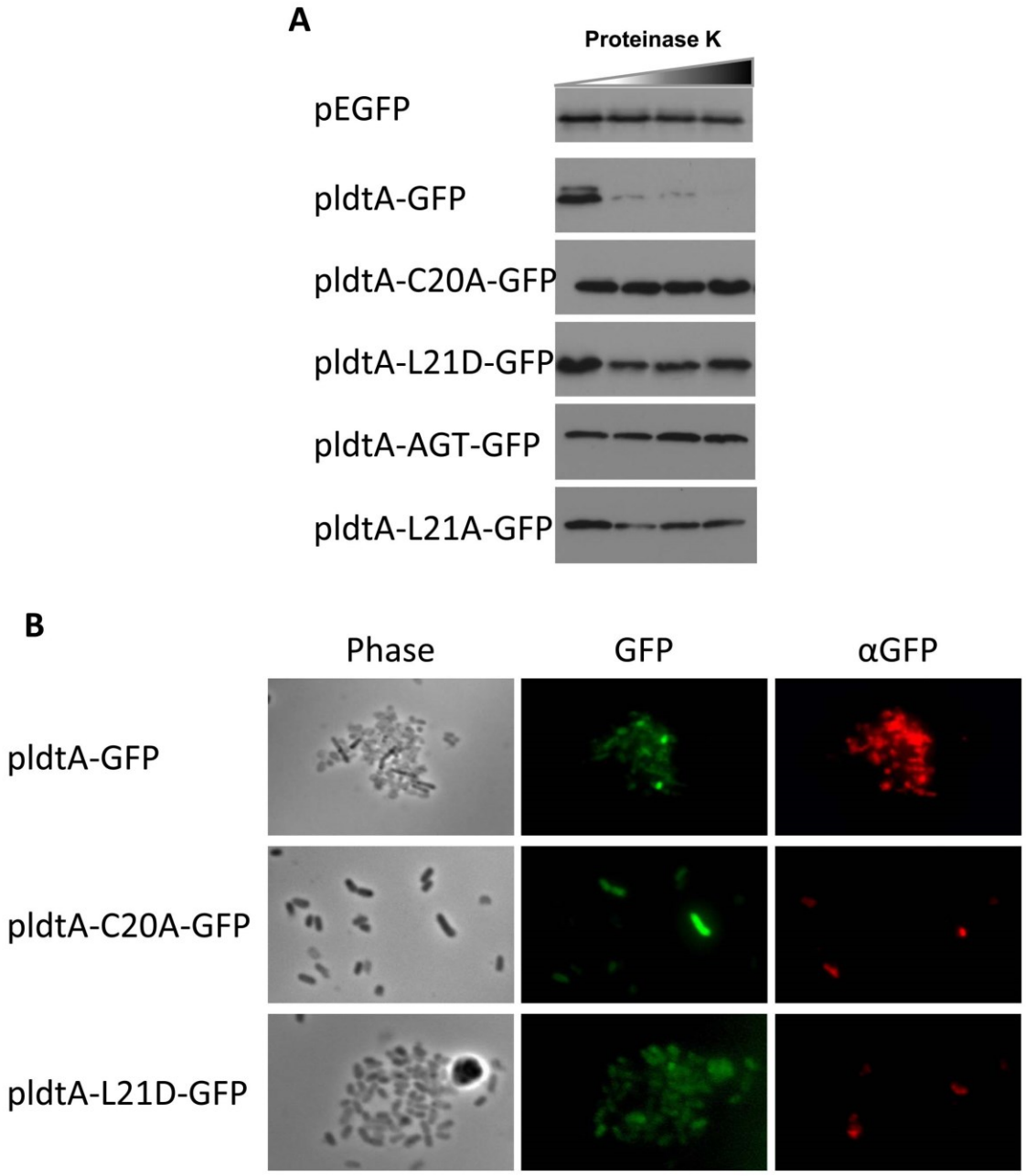

A
Proteinase K
pEGFP
pldtA-GFP
pldtA-C20A-GFP
pldtA-L21D-GFP
pldtA-AGT-GFP
pldtA-L21A-GFP
B
Phase
GFP
αGFP
pldtA-GFP
pldtA-C20A-GFP
pldtA-L21D-GFP

Figure 4.

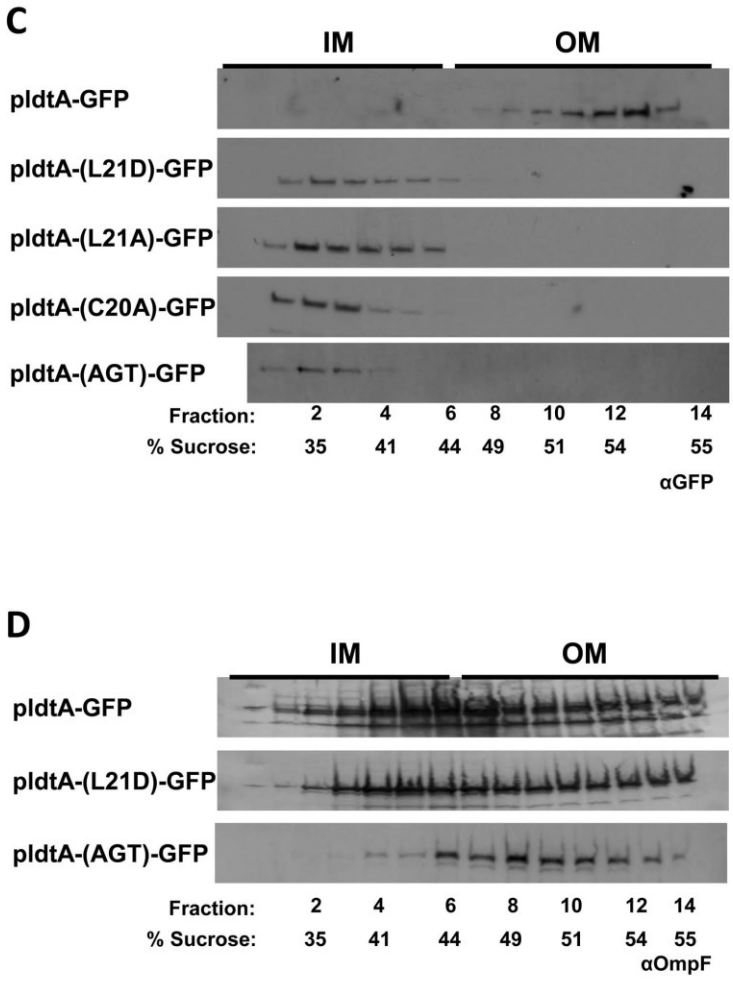

C
IM
OM
pldtA-GFP
pldtA-(L21D)-GFP
pldtA-(L21A)-GFP
pldtA-(C20A)-GFP
pldtA-(AGT)-GFP
Fraction: 2 4 6 8 10 12 14
% Sucrose: 35 41 44 49 51 54 55
αGFP
D
IM
OM
pldtA-GFP
pldtA-(L21D)-GFP
pldtA-(AGT)-GFP
Fraction: 2 4 6 8 10 12 14
% Sucrose: 35 41 44 49 51 54 55
αOmpF

Figure 5.

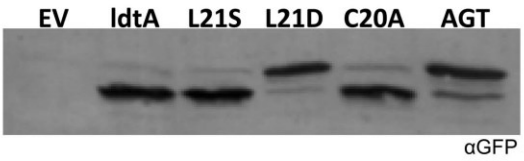


Figure 6.

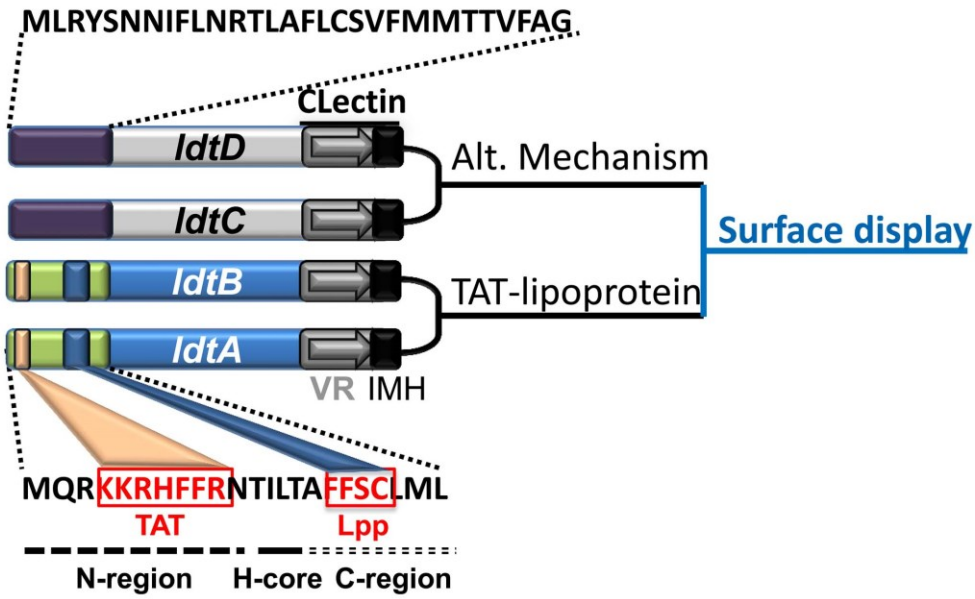
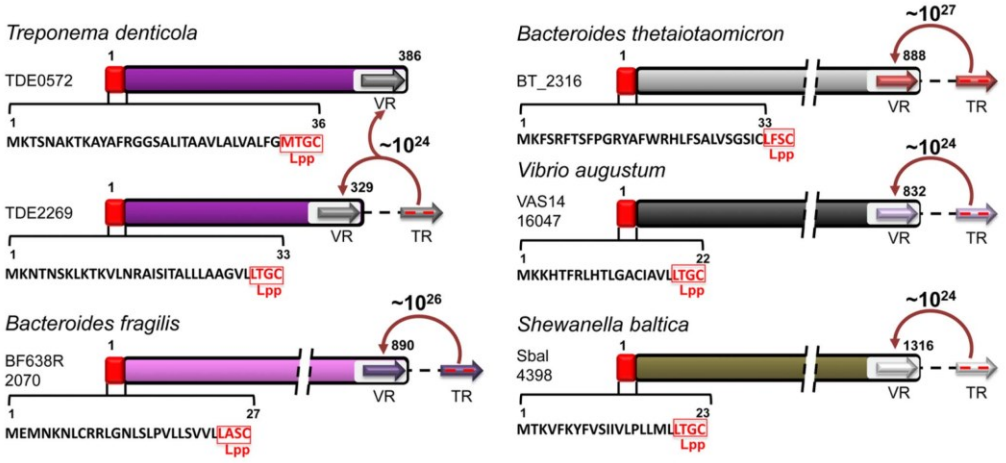
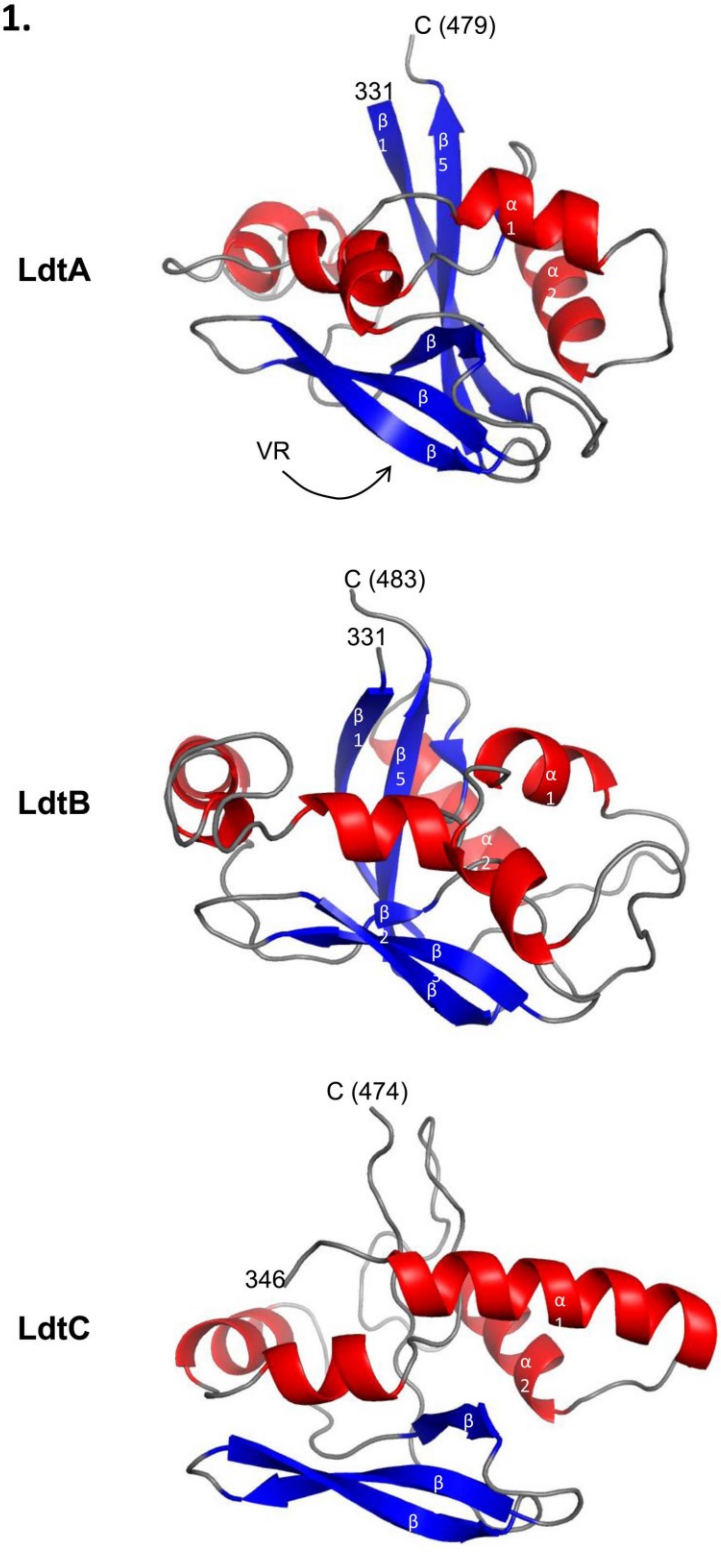

MLRYSNNIFLNRTLAFLCSVFMMTTVFAG
CLectin
IdtD
IdtC
IdtB
IdtA
Alt. Mechanism
TAT-lipoprotein
Surface display
VR IMH
MQRKKRHFFRNTILTAFFSCLML
TAT
Lpp
N-region
H-core
C-region

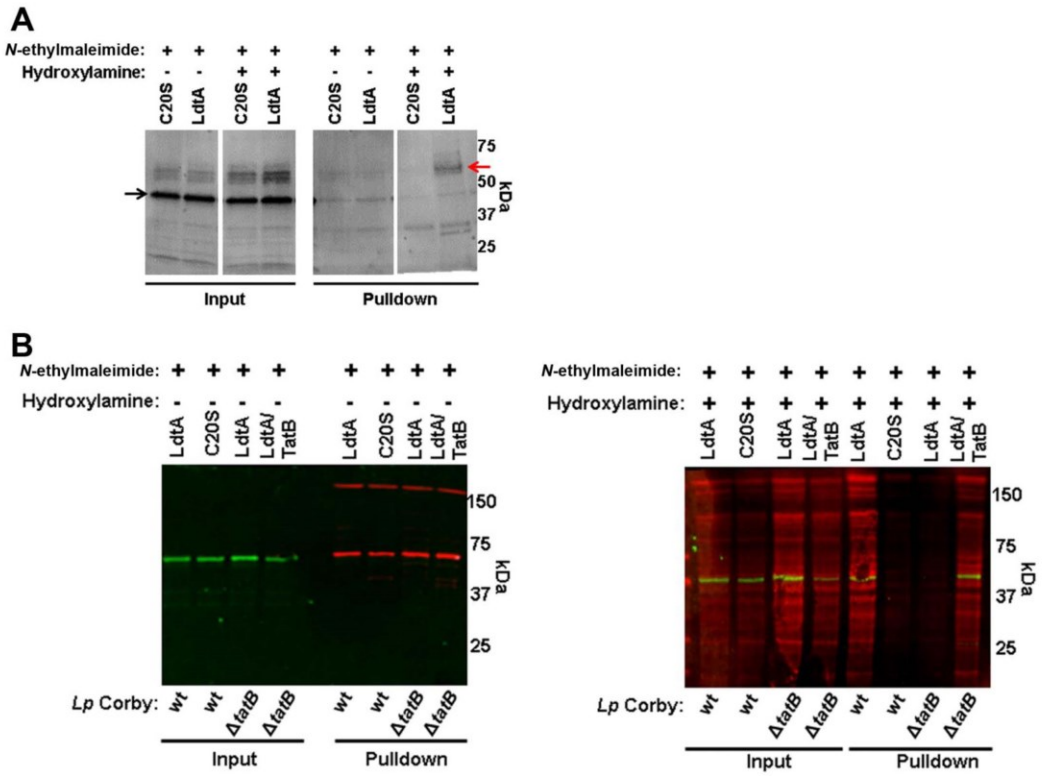
Figure 7.


Treponema denticola
TDE0572
1
386
VR
1
36
MKTSNAKTKAYAFRGGSALITAAVLALVALFGMTGC
Lpp
~10^24
TDE2269
1
329
VR
TR
1
33
MKNTNSKLKTKVLNRAISITALLLAAGVLLTGC
Lpp
Bacteroides fragilis
~10^26
BF638R
2070
1
890
VR
TR
1
27
MEMNKNLCRRLGNLSLPVLLSVVLLASC
Lpp
Bacteroides thetaiotaomicron
~10^27
BT_2316
1
888
VR
TR
1
33
MKFSRFTSFPGRYAFWRHLFSALVSGSICLFSC
Lpp
Vibrio augustum
~10^24
VAS14
16047
1
832
VR
TR
1
22
MKKHTFRLHTLGACIAVLLTGC
Lpp
Shewanella baltica
~10^24
Sbal
4398
1
1316
VR
TR
1
23
MTKVFKYFVSIIVLPLLMLLTGC
Lpp

Sup. Figure 1.


LdtA
C (479)
331
β 1
β 5
α 1
α 2
β
β
β
VR
LdtB
C (483)
331
β 1
β 5
α 1
α 2
β 2
β
β
LdtC
C (474)
346
α 1
α 2
β
β
β

Sup. Figure 2.


A
N-ethylmaleimide: + + + +
Hydroxylamine: - - + +
C20S LdtA C20S LdtA
75
50
37
25
kDa
Input
Pulldown
B
N-ethylmaleimide: + + + + + + + +
Hydroxylamine: - - - - + + + +
LdtA C20S LdtA LdtA/ TatB
150
75
37
25
kDa
Lp Corby: wt wt ΔtatB ΔtatB
Input
Pulldown

References

1. Gogvadze, E. and A. Buzdin, *Retroelements and their impact on genome evolution and functioning.* Cell Mol Life Sci, 2009. **66**(23): p. 3727-42.
2. Liu, M., et al., *Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.* Science, 2002. **295**(5562): p. 2091-4.
3. Doulatov, S., et al., *Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements.* Nature, 2004. **431**(7007): p. 476-81.
4. Medhekar, B. and J.F. Miller, *Diversity-generating retroelements.* Curr Opin Microbiol, 2007. **10**(4): p. 388-95.
5. Mattoo, S., et al., *Mechanisms of Bordetella pathogenesis.* Front Biosci, 2001. **6**: p. E168-86.
6. Guo, H., et al., *Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification.* Mol Cell, 2008. **31**(6): p. 813-23.
7. Guo, H., et al., *Target site recognition by a diversity-generating retroelement.* PLoS Genet, 2011. **7**(12): p. e1002414.
8. Miller, J.L., et al., *Selective ligand recognition by a diversity-generating retroelement variable protein.* PLoS Biol, 2008. **6**(6): p. e131.
9. McMahon, S.A., et al., *The C-type lectin fold as an evolutionary solution for massive sequence variation.* Nat Struct Mol Biol, 2005. **12**(10): p. 886-92.
10. Le Coq, J. and P. Ghosh, *Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement.* Proc Natl Acad Sci U S A, 2011. **108**(35): p. 14649-53.
11. Schillinger, T., et al., *Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF.* BMC Genomics, 2012. **13**: p. 430.
12. Minot, S., et al., *Hypervariable loci in the human gut virome.* Proc Natl Acad Sci U S A, 2012. **109**(10): p. 3962-6.
13. Shin, S., *Innate Immunity to Intracellular Pathogens: Lessons Learned from Legionella pneumophila.* Adv Appl Microbiol, 2012. **79**: p. 43-71.
14. Isberg, R.R., T.J. O'Connor, and M. Heidtman, *The Legionella pneumophila replication vacuole: making a cosy niche inside host cells.* Nat Rev Microbiol, 2009. **7**(1): p. 13-24.
15. Ninio, S., J. Celli, and C.R. Roy, *A Legionella pneumophila effector protein encoded in a region of genomic plasticity binds to Dot/Icm-modified vacuoles.* PLoS Pathog, 2009. **5**(1): p. e1000278.
16. Langille, M.G., W.W. Hsiao, and F.S. Brinkman, *Detecting genomic islands using bioinformatics approaches.* Nat Rev Microbiol, 2010. **8**(5): p. 373-82.
17. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2010. **38**(Database issue): p. D5-16.
18. Stanley, N.R., T. Palmer, and B.C. Berks, *The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in Escherichia coli.* J Biol Chem, 2000. **275**(16): p. 11591-6.

19. Robinson, C., et al., *Transport and proofreading of proteins by the twin-arginine translocation (Tat) system in bacteria.* Biochim Biophys Acta, 2011. **1808**(3): p. 876-84.
20. Okuda, S. and H. Tokuda, *Lipoprotein sorting in bacteria.* Annu Rev Microbiol, 2011. **65**: p. 239-59.
21. Kamalakkannan, S., et al., *Bacterial lipid modification of proteins for novel protein engineering applications.* Protein Eng Des Sel, 2004. **17**(10): p. 721-9.
22. De Buck, E., et al., *A putative twin-arginine translocation pathway in Legionella pneumophila.* Biochem Biophys Res Commun, 2004. **317**(2): p. 654-61.
23. Narita, S. and H. Tokuda, *Sorting of bacterial lipoproteins to the outer membrane by the Lol system.* Methods Mol Biol, 2010. **619**: p. 117-29.
24. Shruthi, H., M.M. Babu, and K. Sankaran, *TAT-pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes.* J Mol Evol, 2010. **70**(4): p. 359-70.
25. Gralnick, J.A., et al., *Extracellular respiration of dimethyl sulfoxide by Shewanella oneidensis strain MR-1.* Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4669-74.
26. Helbig, J.H., et al., *Immunolocalization of the Mip protein of intracellularly and extracellularly grown Legionella pneumophila.* Lett Appl Microbiol, 2001. **32**(2): p. 83-8.
27. Roy, C.R. and R.R. Isberg, *Topology of Legionella pneumophila DotA: an inner membrane protein required for replication in macrophages.* Infect Immun, 1997. **65**(2): p. 571-8.
28. Miyamoto, H., et al., *Protein profiles of Legionella pneumophila Philadelphia-1 grown in macrophages and characterization of a gene encoding a novel 24 kDa Legionella protein.* Microb Pathog, 1993. **15**(6): p. 469-84.
29. Matthews, M. and C.R. Roy, *Identification and subcellular localization of the Legionella pneumophila IcmX protein: a factor essential for establishment of a replicative organelle in eukaryotic host cells.* Infect Immun, 2000. **68**(7): p. 3971-82.
30. Schulze, R.J. and W.R. Zuckert, *Borrelia burgdorferi lipoproteins are secreted to the outer surface by default.* Mol Microbiol, 2006. **59**(5): p. 1473-84.
31. De Buck, E., et al., *Legionella pneumophila Philadelphia-1 tatB and tatC affect intracellular replication and biofilm formation.* Biochem Biophys Res Commun, 2005. **331**(4): p. 1413-20.
32. Rossier, O. and N.P. Cianciotto, *The Legionella pneumophila tatB gene facilitates secretion of phospholipase C, growth under iron-limiting conditions, and intracellular infection.* Infect Immun, 2005. **73**(4): p. 2020-32.
33. Linder, M.E. and R.J. Deschenes, *Palmitoylation: policing protein stability and traffic.* Nat Rev Mol Cell Biol, 2007. **8**(1): p. 74-84.
34. Morales, V.M., A. Backman, and M. Bagdasarian, *A series of wide-host-range low-copy-number vectors that allow direct screening for recombinants.* Gene, 1991. **97**(1): p. 39-47.
35. Hara, T., S. Matsuyama, and H. Tokuda, *Mechanism underlying the inner membrane retention of Escherichia coli lipoproteins caused by Lol avoidance signals.* J Biol Chem, 2003. **278**(41): p. 40408-14.
36. Cowles, C.E., et al., *The free and bound forms of Lpp occupy distinct subcellular locations in Escherichia coli.* Mol Microbiol, 2011. **79**(5): p. 1168-81.

37. Seydel, A., P. Gounon, and A.P. Pugsley, *Testing the '+2 rule' for lipoprotein sorting in the Escherichia coli cell envelope with a new genetic selection.* Mol Microbiol, 1999. **34**(4): p. 810-21.
38. Chatzi, K.E., et al., *Breaking on through to the other side: protein export through the bacterial Sec system.* Biochem J, 2013. **449**(1): p. 25-37.
39. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.
40. Juncker, A.S., et al., *Prediction of lipoprotein signal peptides in Gram-negative bacteria.* Protein Sci, 2003. **12**(8): p. 1652-62.
41. Bendtsen, J.D., et al., *Prediction of twin-arginine signal peptides.* BMC Bioinformatics, 2005. **6**: p. 167.
42. Shruthi, H., et al., *Twin arginine translocase pathway and fast-folding lipoprotein biosynthesis in E. coli: interesting implications and applications.* Mol Biosyst, 2010. **6**(6): p. 999-1007.
43. Paetzel, M., et al., *Signal peptidases.* Chem Rev, 2002. **102**(12): p. 4549-80.
44. Zelensky, A.N. and J.E. Gready, *The C-type lectin-like domain superfamily.* FEBS J, 2005. **272**(24): p. 6179-217.
45. Lautner, M., et al., *Regulation, integrase-dependent excision, and horizontal transfer of genomic islands in Legionella pneumophila.* J Bacteriol, 2013. **195**(7): p. 1583-97.
46. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
47. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
48. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. **4**(3): p. 363-71.
49. Batchelor, M., et al., *Structural basis for recognition of the translocated intimin receptor (Tir) by intimin from enteropathogenic Escherichia coli.* EMBO J, 2000. **19**(11): p. 2452-64.

Chapter 5. Future research and perspectives.

Summary

Diversity-generating retroelements (DGRs) are a relatively recent discovery, initially described in 2002 [1], and thought to be novel among retroelements because of their ability to rapidly evolve the ligand binding domains of target proteins (TPs) and the derived benefit diversified TPs convey to their hosts. To date, the archetype DGR, the *Bordetella bronchiseptica* bacteriophage (BPP), has been the sole paradigm for mechanistic studies of mutagenic homing as well as the initial structural studies which demonstrated the exquisite co-evolution of a system that generates nucleotide diversity and a protein scaffold, the C-type lectin scaffold (CLec), which balances the required flexibility of displaying diversity within a static scaffold [1-3]. This balance ensures TP-ligand interactions occur in a contextually meaningful manner.

Attempts to enumerate the breadth, in both total numbers and range of hosts, of DGRs within the three domains of life identified over three-hundred potential elements in at least 20 phyla of bacteria and one phylum of archaea [4, 5]. This is likely an underestimation due to the constant addition of sequences, representing new strains and species, to nucleotide sequence databases. As a family, DGRs are found in plasmids, phage, and chromosomes, their inheritance can be vertical or horizontal, and their targets proteins are as diverse as the hosts in which they are found.

This leads to some very basic questions about their biology. Does adenine mutagenesis occur through the same mechanism? As diverse as TPs are, what biological roles do they play in their hosts? What determines the intra- and inter-species distribution of DGRs? Why are some retroelements beneficial to their hosts while others

are overtly deleterious? DGRs provide an opportunity to study a class of mobile genetic elements in hosts of nearly every description but model systems are clearly needed. As described in the preceding chapters, we have identified a putative DGR within the opportunistic, human pathogen *Legionella pneumophila* (*Lp*). We analyzed the genetic components of the diversification machinery and determined their contribution to mutagenic homing, we characterized the TP to understand how and where *Lp* displays a massively variable protein and finally, we probed the breath of DGRs within the *Legionella* genus, discovering that the same element has been widely distributed. We have demonstrated that *Lp* is a model system for analysis of the mechanism of DGR mutagenic homing and its genetic tractability makes it a candidate for dissecting the pathways by which proteins are displayed upon the bacterial surface. Furthermore, *Legionella* clinical and environmental isolates from worldwide sources have been extensively studied as well as organized into data sets which are amenable to investigating non-random associations between DGRs and strain source, genotype, or serotype. Using this system we propose:

1. To investigate the regulation of DGRs by *Lp* global signal cascades. We have demonstrated increased expression of TR-RNA in *relA*/*spoT* mutant cells that coincides with drastically increased levels of target gene diversification. Using a combination of gene and protein expression studies we will address the hypothesis that DGRs are being regulated by host networks known to function in phenotypic variation and respond to environmental stress. We will probe the exact regulators as well as branches used by *Lp* to generate diversity within the target gene. This analysis could provide a dichotomy

to the archetype element BPP where diversification is thought to occur at a stochastic rate.

2. DGRs have coevolved a system to generate diversity in target genes with a protein scaffold to display variability. We will continue our analysis using an unbiased screen and directed deletion of key pathways to understand the precise mechanisms used to display a lipoprotein upon the surface of gram negative bacteria. Surface display of TPs appears to be a common feature of many DGRs and our preliminary data suggests it may require host systems common to all bacteria.

3. We will take advantage of the wide spread study of *Legionella* genomes to understand the distribution of DGRs throughout this genus. This work could provide insight into if these elements are found preferentially within clinical, environmental, or serogroup-specific isolates.

The regulation of DGRs within Legionella.

Background

Our investigation into the *Lp* strain Corby DGR started from a bioinformatics screen of sequenced nucleotide databases [6]. The *Lp* DGR is striking because of its potential for diversification of *ldtA*, the 43 adenines in TR equates to roughly 10^{23} nucleotide permutations within VR, or upon translation ~10^{19} polypeptide sequences within the C-terminal CLec domain of LdtA. For comparison, the mammalian humoral immune system has been postulated to be capable of generating ~10^{16} nucleotide

permutations within the hypervariable loop of the antibody scaffold [7, 8] and practical applications like phage display of antibody libraries typically generate libraries of ~10^6 to 10^{10} [9, 10] making the *Lp* DGR the largest natural diversification system described to date.

We demonstrated the Corby element was active, under *in vitro* endogenous growth conditions, using a plasmid overexpression system, as well as under conditions which mimic the *Lp* biphasic lifecycle, through tagging of the chromosomal TR with an invariant sequence and monitoring its transfer to VR. Interestingly, transfer of the invariant tag into VR was most readily detected in cells grown to stationary phase, which is typically when *Lp* transitions from the replicative to transmissive state [11] and suggested that the enzymes *relA* and *spoT*, which coordinate this shift in phenotypic state, may also be involved in regulation of DGR genes. Analysis of *Lp* mutants, where both *relA* and *spoT* had been deleted, demonstrated an increase in mutagenic homing that began during the transition from replicative to transmissive state, but was most evident in cells completely modulated to the transmissive phase. Further analysis revealed these mutants displayed levels of mutagenic homing, specifically adenine mutagenesis, similar to overexpression conditions, in contrast to *wt* cells grown to stationary phase where mutagenic homing displays little adenine mutagenesis. Furthermore, the number of TR-RNA transcripts appears to increase in mutant cells as compared to *wt* cells in stationary phase. These data suggest the *Lp* DGR, or at least certain components are regulated and have likely integrated into host global regulatory cascades and that mutagenic homing, specifically adenine mutagenesis, does not occur at the stochastic rates which are thought to result in tropism switching in BPP [1].

The *Lp* RelA and SpoT regulatory cascade. RelA and SpoT regulated gene expression is a highly complex system which has been studied in several bacteria and is generally thought to occur through direct or indirect means. Direct regulation is mediated by a poorly understood mechanism that involves RelA and SpoT synthesized ppGpp as well as the polymerase binding protein DksA interacting with the RNA polymerase and the resulting expression or repression of genes by this complex being dependent upon their promoter sequence [11]. In *Lp*, one direct target is a two component regulatory system (TCRS), LetAS, which senses cellular levels of ppGpp and activates genes through binding of the transcriptional activator LetA to target promoter sequences. Targets of LetA are two regulatory RNAs, *rsmYZ*, which bind to and antagonize CsrA in order to relieve its transcriptional repression of transmissive state genes, e.g. the Dot/Icm type 4 secretion system (T4SS) genes [11]. Indirect gene regulation occurs through sigma factor competition, where increasing levels of ppGpp causes an inhibition of σ^{70} (housekeeping)-driven promoters, allowing for the differential expression of genes using alternative sigma factors [11]. This has been demonstrated for several promoters using the stationary phase (σ^{S}), flagellar (σ^{28}), or alternative (σ^{54}) sigma factors [11] and is essential for *Lp* to transition into the transmissive state. However, regulation in *Lp* is complicated by the observation that many regulators also display ppGpp independent activity. DksA functions in concert with ppGpp during stationary phase, but also functions independently of ppGpp to express a subset of genes during exponential growth phase [11]. The biphasic lifecycle of *Lp* is a highly coordinated series of senses and responses driven, in most part, by the activity of RelA and SpoT.

In addition to the direct and indirect regulation of genes, ppGpp modulate several other bacterial processes which could have an effect on DGR activity. Intracellular levels of ppGpp have been shown to affect mRNA half-life in several organisms, with *relA*/*spoT* mutants experiencing shorter mRNA half-lives than *wt* cells. This is thought to occur through the direct modulation of enzymes, like polynucleotide phosphorylase, which facilitates mRNA degradation or *pcnB*, which polyadenylates transcripts, thus increasing their stability [12]. In addition to affecting mRNA stability, ppGpp binding has been shown to affect the activity of enzymes like CadA, a member of the lysine-cadaverine antiport system that facilitates bacterial survival in acidic environments [12]. Levels of the signaling molecule polyphosphate also appear to be partially regulated by intracellular stores of ppGpp. Polyphosphate plays a role in global stress response by inhibiting the expression of ribosomal gene but it has also been implicated in the expression of virulence traits [12, 13].

Other *Lp* regulatory cascades. *RelA* and *spoT* play a pivotal role in gene regulation and the progression of *Lp* through its biphasic lifecycle however, there are other independently regulated signal cascades and recent evidence suggests cross talk between these systems. The Dot/Icm T4SS is essential for delivery of protein effectors into host cells and results in modification of the endosome into the replication competent legionella containing vacuole. Several key effectors have been demonstrated to be solely regulated late in the transmissive state by the response regulator, PmrA, and this is thought to prime the bacterium for host cell invasion [14]. While the environmental trigger for PmrA has not been identified, it has recently been shown that its regulation of a subset of genes is affected by CsrA, as well as the non-coding RNAs RsmZY, which

are directly controlled by LetAS and in turn by levels of ppGpp. This suggests that *Lp* genes can be regulated transcriptionally and/or post-transcriptionally by multiple networks and this is likely necessary for the fine tuning of responses [14, 15].

Investigating the effect of key regulators on levels of DGR mutagenic homing.

Our analysis of the regulation of DGRs in *Lp* demonstrated that deletion of *relA* and *spoT* resulted in increased levels of mutagenic homing. *RelA* and *spoT* were initially chosen for three reasons: they have been shown to affect levels of retrohoming for the related group II intron LI.LtrB [16], they are directly responsible for the differential regulation of ~50% of all *Lp* genes and, they coordinate the transition from replicative to transmissive state [17]. We will take advantage of the genetic tractability of *Lp* to generate non-polar deletions, see chapter 2 for a detailed description, of key regulators within the *relA*/*spoT* cascade as well as cascades independently coordinated by other regulators. Using *Lp* Corby cells with an invariant tag inserted into TR (TRwTAG) we will delete the following regulators: the polymerase binding protein DksA, the activator from the LetAS TCRS, and the RNA binding transcriptional repressor CsrA [11, 18]. These mutants will be grown in rich media which duplicates the *Lp* biphasic lifecycle, samples will be taken at time points which correspond to changes in phenotypic state, and the resulting genomic DNA will be assayed for the transfer of the invariant tag from TR to VR by polymerase chain reaction (PCR). This analysis will allow us to address several key questions regarding regulation of DGRs in *Lp*. Is TR directly or indirectly regulated by ppGpp? Does differential regulation of TR occur at the level of LetA (positively) or

CsrA (negatively)? Furthermore, analysis of alternative sigma factors directly regulated by intracellular levels of ppGpp, such as σ^S or σ^{54}, could provide insight into specific cascades necessary for the regulation of DGRs. Our analysis will also include sequencing of PCR products from homing assay for several key reasons. It allows verification of adenine mutagenesis, a crucial step in identifying genuine mutagenic homing, but it will also provide a metric for assessing relative levels of adenine mutagenesis. This would allow us to separate regulators which show increased levels of mutagenic homing from those that show increases in homing but not mutagenesis.

The observation that *relA*/*spoT* mutants only display increased levels of mutagenic homing during the transition to or while completely within the transmissive state suggests that levels of ppGpp only partially regulate DGR activity. Therefore, analysis will be performed (as above) of regulatory branches which function independently of RelA and SpoT. One such target would be the deletion of the TCRS *pmrAB*, based on the observation that PmrA is activated independently of intracellular levels of ppGpp and regulates a subset of genes during late transmissive phase, yet gene regulation by PmrA can involve the LetAS modulated transcriptional repressor CsrA [11, 15]. While PmrAB is an obvious system to investigate there are alternative regulatory systems, like the TCRS CpxAR, which are also though to function independently of RelA/SpoT and could play a role in modulating DGR activity.

Analysis of regulatory mutants using QPCR. DGR mutagenic homing is expected to be highly complex and tightly regulated as the introduction of mutations into coding sequences is generally thought to be deleterious [19]. We hypothesize that DGR

regulation likely bifurcates between expression of the TP and the diversification machinery e.g. *avd*, TR, and *RT*. Currently our observations are limited to the *relA*/*spoT* mutants which show an increase in TR-RNA as compared to *wt* cells, with fairly constant transcript levels of *ldtA* (TP gene), *avd* (data not shown) and *RT* observed in both *wt* and *relA*/*spoT* mutant cells. However, to better understand the regulation of DGRs, we will analyze the changes in *TP*, *avd*, *TR*, and *RT* transcript number by quantitative PCR (qPCR) in all regulatory mutants (see above) which demonstrate altered levels of mutagenic homing, as determined by our PCR assay.

Post-translational regulation. While we have identified and expect to identify additional regulatory branches which have an effect on levels of DGR mutagenic homing, there is a possibility that regulation occurs post-translationally. To address this, we will take advantage of any regulatory mutants we identify and individually replace the chromosomal allele of DGR genes (*ldtA*, *avd*, and *RT*) with epitope tagged variants. Protein levels will be assessed by western blot during the transitions between phenotypic states (see above), allowing the monitoring of protein levels during states where increases in mutagenic homing are observed. There are several possibilities where post-translational regulation would be consistent with the observations regarding transcript number of *avd*, *TR,* and *RT* observed in *relA*/*spoT* mutant cells. Levels of Avd and RT may be consistent, only requiring increased numbers of TR-RNA to complete mutagenic homing. Alternatively, protein levels of DGR genes may fluctuate, for any number of reasons, as *Lp* progresses through its lifecycle.

The experiments outlined above build upon work described in previous chapters. They concisely attempt to identify additional regulatory braches which might affect levels of DGR mutagenic homing through targeted deletion of key regulators which are responsible for modulating individual pathways in *Lp*. These experiments could provide insight to how a horizontally acquired DGR, and perhaps other mobile retroelements, integrate in host global networks and further our understanding of the basic biology of DGRs.

Bacterial surface display of TAT secreted lipoproteins.

Background

Lipoproteins function in a wide variety of cellular processes and since their identification have generally been considered to be translocated across the gram negative inner membrane by the general secretory pathways (SEC) where they are modified, retained in the inner membrane, or translocated across the periplasm to the outer membrane by localization of lipoprotein (LOL) system [20]. The twin-arginine translocation (TAT) system is an alternative to SEC for translocating folded proteins or protein complexes across lipid bilayers [21]. Although TAT secreted lipoproteins, as a distinct class of lipoproteins, were described relatively recently they have been identified in at least 696 prokaryotic genomes [22]. TAT-lipoproteins appear to be enriched in extremophiles where the fast folding and stable properties of TAT proteins could provide a selective advantage to their host [22]. Functional analysis of open reading frames containing TAT-lipoprotein motifs found they encode for proteins with predicted

enzymatic, transport, or general binding domains [22, 23]. Several TAT-lipoproteins were identified in *Streptomyces coelicolor* and analysis revealed their transport dependent upon *tatC* [24]. Maturation of several TAT-lipoproteins found in *Haloferax volcanii* depended on residues within the TAT motif, the lipobox conserved cysteine, and was inhibited by globomycin, an antagonist of signal peptidase II [25]. These experiments analyzed the requirement of conserved residues within protein trafficking motifs recognized by proteinaceous components of secretion systems responsible for transporting substrates. This work suggests that the swapping the SEC motif with a TAT motif is a functionally analogous replacement and TAT-lipoproteins are produced, modified, and translocated similarly to lipoproteins transported by the SEC pathway.

DGR mutagenic homing has coevolved along with the C-terminal CLec domain which displays diversified residues and this system is thematically similar to the coevolution of diversity displayed by the antibody scaffold by the mammalian immune system. However, the diversified CLec domain must be displayed in a meaningful context upon the surface of a bacterium. We analyzed the signal peptide of *ldtA* to better understand the context in which *Lp* displays a massively variable protein upon its cell surface.

TAT-lipoprotein LdtA traffics to the cell surface by an unusual mechanism. *In silico* analysis of the *Lp* Corby DGR target gene, *ldtA*, identified a bipartite signal peptide composed of a TAT motif followed by a lipobox motif. As discussed in greater detail in chapter four, we hypothesized that LdtA may be localized in the outer leaflet of the *Lp* outer membrane and this was confirmed using a combination of molecular

biology and microscopy. We then sought to determine the precise pathways involved in the surface display of TAT-lipoproteins. The requirement of the TAT system for translocation of LdtA across the bacterial inner membrane was demonstrated through deletion of *tatB* [26]. The identification of a lipobox within LdtA made periplasmic recognition followed by outer membrane localization via LOL likely since it is the only known system capable of transporting lipoproteins to the bacterial outer membrane [27]. Lipoproteins, regardless of the pathway used to cross the inner membrane, are modified sequentially by three well conserved enzymes which typically results in cleavage of its signal peptide and acylation of the lipobox's conserved cysteine [27]. Cleavage and modification of the conserved cysteine are crucial for biogenesis of lipoproteins as these steps are thought necessary for their insertion and retention in lipid bilayers. We demonstrated that replacement of the LdtA conserved cysteine inhibits its surface display and used acyl biotin-exchange chemistry to demonstrate the post-translational modification of the conserved cysteine [28]. The residues, called targeting residues, following the lipoprotein conserved cysteine are crucial in determining to which membrane lipoproteins would be sorted to via LOL [27, 29]. Substitution of targeting residues in LdtA with a range of amino acids, even chemically similar ones, resulted in retention of LdtA in the inner membrane. Furthermore, cleavage of the LdtA signal peptide appears to depend on the residue at the +2 position and, to our knowledge, this observation has not been reported [30]. These data suggest that LdtA, and perhaps all TAT secreted lipoproteins, is trafficked to the cell surface by a poorly understood mechanism that differs from that used to traffic lipoproteins to the periplasmic face of the bacterial outer membrane.

Identifying host systems necessary for the surface display of LdtA.

We have demonstrated the requirement of the TAT system for translocation across the *Lp* inner membrane. It is at this inner membrane/periplasmic interface where LdtA is likely cleaved and modified by Lgt/LspA/Lnt into a mature lipoprotein. This matured peptide would be recognized by the Lol system for transport to the outer membrane where an unknown factor, likely a “flipase”, resolves it to the environmental facing facet of the outer membrane. We propose to target these two pathways to determine the precise mechanism by which LdtA is transported from cytoplasm to surface using a two pronged approach. A screen to identify host factors necessary for surface display and in conjunction, targeted deletions of key proteins within specific pathways.

A screen for host factors necessary for surface display of LdtA. As described in greater detail in chapter four, we will take advantage of the fact that fusing the LdtA signal peptide [26] to a protein of interest is sufficient for surface display in both *Lp* and *E. coli*, this was demonstrated for both green fluorescent protein (GFP) and β-Lactamase (data not shown). We will construct a plasmid, in wild type *E. coli* strain MG1655, where the LdtA signal sequence is fused in-frame with modified GFP followed by an additional peptide tag which is derived from the *E. coli* acyl carrier protein (ACP). The ACP-tag is a small epitope tag which can be modified through the action of 4’-phosphopantetheinyl transferase (SFP) with derivatives of coenzyme A (CoA) which are themselves modified with fluorophores. This system is commercially available and has

recently been shown capable of specifically labelling an outer membrane porin, LamB, of *E. coli* [31].

By fusing the LdtA signal sequence with GFP and ACP-tag, a dual epitope tagged version of LdtA is generated which allows for differential detection of surface exposed, as compared to intra-cellular proteins, based on detection of a single/dual fluorescence emission. It has been demonstrated that TAT secreted, periplasmic GFP folds properly and fluoresces [32]. LdtA-GFP-ACP-tag that is trafficked to the cell surface retains its GFP fluorescence, but now the ACP-tag can be ligated with CoA-547, a non-cell permeable fluorescent substrate (New England Biolabs). Cells will be analyzed and separated by fluorescence-activated cell sorting (FACS) based on cells emission at 488 nm (GFP/green) and 568 nm (CoA-547/red) as an indicator of sub-cellular (green fluorescence only) vs. sub-cellular and surface exposed (green and red fluorescence) LdtA. We will use a plasmid based vector pMMB208 [33] to express *ldtA-GFP-ACP-tag* under the control of the *tac* promoter, generating p208-ldtA-GFP-ACP. Labelling with CoA-547 is not lethal nor does it require fixation so cells can be treated to multiple rounds of selection and separation, with individual cells being recovered for further analysis. Controls include LdtA signal peptides with *wt*, mutation of the conserved cysteine (C20A), and mutation of the TAT-motif conserved arginine (R6A) sequences which have been shown to be trafficked to the cell surface, periplasm, or cytoplasm, respectively [25, 26].

Screening an *E. coli* knockout library. To screen *E. coli* for host factors necessary for surface display of LdtA we will take advantage of libraries containing

precisely defined single-gene deletions, such as the Keio collection. This commercially available collection of *E. coli* mutants is constructed in a K-12 derivative, similar to MG1655, and represents a deletion of all non-essential genes, in duplicate [34]. Cells will be transformed with p208-ldtA-GFP-ACP, grown in rich media, and induced for protein expression and then analyzed by FACS, being separated into two populations based on the detection green +/- red fluorescence. Simultaneous detection of green and red fluorescence in *E. coli* has been previously demonstrated as a robust, simple detection method [35]. Cells recovered with strong green fluorescence can be treated to serial rounds of induction and analysis to ensure loss of CoA-547 is not due to technical issues.

The above experiments build upon observations made in this lab and we expect to identify known, as well as novel, genes involved in surface display of LdtA. Our assay is high-throughput, since it does not require screening of individual plates/colonies and robust because of the ability to iteratively analyze putative clones.

Contribution of the LOL system to surface display of LdtA.

The LOL system is thought to recognize matured lipoproteins by the chemical characteristic of the conserved cysteine/targeting residues by the trinary inner membrane complex, LolCDE. This complex shares similarity with ATP-binding cassette transporters and is composed of two membrane spanning proteins, LolCE, and an ATPase, LolD. After recognition of a lipoprotein destined for the outer membrane, ATP hydrolysis occurs inducing a conformational shift in the LolCDE complex and the protein

is transferred to the periplasmic chaperone, LolA. The lipoprotein is ferried across the periplasm where an affinity based interaction transfers it from LolA to the outer membrane localized LolB which inserts the protein in the outer membrane lipid bilayer [27].

Generating a conditional lethal for *lolA*. Although the Lol system is necessary for life, we propose to determine its contribution to surface display of LdtA using a conditional lethal system. We will construct a plasmid based system to ectopically express epitope tagged *lolA* in *wt E. coli* strain MG1655 under the control of the titratable promoter pAraBAD (pBAD-lolAHIS) [36]. Additionally, we will construct an allelic exchange deletion vector for *lolA* using the sucrose counter-selectable suicide plasmid pRE118, generating pRE118ΔlolA [37]. MG1655 cells with be transformed with pBAD-lolAHIS, transformants mated with the *E. coli* auxotroph helper strain RHO3 carrying pRE118ΔlolA. The mating will be plated on LB + 0.2% Arabinose and diaminopimelic acid (DAP) in order to induce for the expression of LolAHIS and allow for growth of RHO3, respectively. Cells will then be streaked onto LB containing 0.2% Arabinose and kanamycin (Km) to allow for expression of LolAHIS and to select for integration of the suicide plasmid into the MG1655 chromosome; removal of DAP will select against RHO3 cells which are Km resistant. Integrants will be plated on LB containing 10% sucrose and 0.2% arabinose to select for resolution and deletion of *lolA*. Cells will be tested for growth only in the presence of arabinose and can be counter-selected in the presence of glucose which represses expression of the AraBAD promoter. We have chosen to generate this system in *E. coli* since a similar strategy

has already been shown to be successful [38] and that there are currently no plasmids with the same dynamic range as pAraBAD in *Lp*.

Assays for surface display of LdtA. Surface display of LdtA will be investigated in MG1655Δ*lolA* + pBAD-lolAHIS using well established assays that we have previously published with [26]. In short, MG1655Δ*lolA* + pBAD-lolAHIS cells will be transformed with pMMB208-ldtA-HA, cells will be maintained in LB with 0.2% arabinose and 1 mM IPTG to induce the expression of LolAHIS and LdtAHA, respectively. Replacement of arabinose with 0.2% glucose will cause the direct repression of pBAD-lolAHIS without affecting expression of LdtAHA. Whole cells +/- treatment with glucose will be subject to immuno-microscopy, protease sensitivity, and membrane fractionation to determine alterations in sub-cellular localization of LdtA, as described in chapter 4. Controls will include the classically studied Braun's lipoprotein (lpp) as a positive control [39] and the beta-barrel protein OmpA as the negative control [40].

The above experiments are outlined to determine the contribution of the Lol system to surface display of LdtA. While we realize these experiments are technically challenging, they would provide insight into a conserved mechanism for displaying lipoproteins upon a bacterial cell surface. Furthermore, understanding the precise pathways necessary to direct a protein to either the cytoplasm, inner membrane, or either facet of the outer membrane would be of great utility and have biotechnological applications.

Identification of the "flipase" required for surface display of LdtA. The factor or system necessary to translocate LdtA from the periplasmic leaflet to the extracellular

facet of the outer membrane is unknown and has been hypothesized as a flipase [23, 27]. While our screen (see above) may identify this factor, we will also pursue a targeted approach. The type 2 secretion system (T2SS) is one of six secretion systems typically found in gram-negative bacteria and is responsible for the release of a large number of proteins from the bacterial periplasm [41]. While T2SSs are generally considered important due to the necessity of the release of soluble proteins for virulence, several recent reports have found them necessary for the release of several lipoproteins. Furthermore, T2SS was shown to be required for the function of a TAT-lipoprotein DmsA, although its precise role is still unclear [23].

Analysis of Lp T2SS mutants. We will generate *Lp* Corby T2SS mutant cells by deleting the inner membrane scaffold gene, *lspF*, using allelic exchange with *sacB* counter selection, a mutation that has previously shown to ablate activity [42, 43]. *LpΔlspF* cells will be transformed with an epitope tagged *ldtA*, expressed from a plasmid vector under the control of the *tac* promoter. Wild type and mutant cells will be induced for protein expression and assayed for surface localization of LdtA using immunofluorescence and treatment with exogenous proteases, assays well described in chapter 4. Controls for localization of LdtA will include *wt* Corby cells as well as Corby*ΔtatB* while controls for T2SS activity will be to probe for CelA, a glycosyl hydrolase, in supernatant vs. cell pellet using a commercially available antibody [42].

Cumulatively, these experiments represent a complementary approach, an unbiased screen for bacterial host factors required for, as well as the direct assessment of the contribution of specific secretion pathways to the surface display of LdtA.

Furthermore, surface display of TAT-lipoproteins appears to require pathways conserved across several species suggesting surface display of lipoproteins may be a common feature of bacteria.

Distribution of DGRs within the *Legionella* genus

Over 84% of legionellosis cases are caused by *Lp* from serogroup 1 (Sg1) [44]. As described in chapter 3, our preliminary survey of 12 Sg1 isolates identified homologous DGRs in three and bioinformatics analysis of whole-genome shotgun contigs identified three additional elements, two in *Lp* and one in *Legionella tunisiensis* [45, 46]. We have proposed collaboration with Dr. Natalia Kozak at the Center for Disease Control and Prevention [47], to conduct a large scale assessment of the distribution of DGRs in *Lp*. We are fortunate to have access to a collection of isolates that have recently been characterized by Dr. Kozak using sequence-based typing (SBT) analysis [47]. The CDC collection includes 540 sporadic clinical isolates, 90 environmental isolates with no known association with disease, and 170 outbreak isolates, all of which are from Sg1. Results from Dr. Kozak's analysis show that a limited number of sequence types (STs) are associated with both a majority of outbreaks and sporadic cases of disease.

Screening a large library of *Legionella* strains. We have established PCR assays for DGR components that detect conserved regions of *avd* as well as *RT*, and these will be used to probe DNA samples available for nearly all of the 800 isolates in the CDC collection. An initial screen will select a subset of ~100 isolates based on

source and phylogenetic distribution. Amplified products of expected size will be sequenced to determine if they are truly DGR-derived and, if so, we will sequence the entire element. We will also determine if putative elements are found within a similar genomic island as known DGRs as well as the relative genomic location by sequencing chromosomal junction fragments. If we suspect a non-random distribution in a particular clonal complex or clade, additional samples will be selected and analyzed to increase resolution.

Bioinformatic and statistical methods for establishing relationships between STs and for inferring phylogenetic relationships and non-random associations with genes and alleles are well described [47]. We will correlate DGR presence, type, and genomic location with phylogenetic relationships between strains, and determine if DGRs are overrepresented in clinical vs. environmental isolates. We will also identify potential associations with STs that are correlated with enhanced virulence. Differences in target gene VR sequences will reveal the extent of DGR activity that exists in nature and similarities or differences in chromosomal location will indicate independent horizontal transfer events vs. ancestral origin.

This analysis will expand on our initial studies into the distribution of DGRs within the *Legionella* genus. Screening of large sample sets will yield a wealth of information regarding the presence of elements within clinical vs. environmental isolates but also the distribution within serogroups which show a range of pathogenicity. This work will provide our first glimpse into the evolutionary dynamics of DGRs in a bacterial species.

Conclusion

Retroelements are found in all domains of life and are often considered, at worst, to be selfish elements or, at best, neutral to their host [48]. In contrast, diversity-generating retroelements have been shown to accelerate the evolution of target protein ligand binding domains and for the bacteriophage BPP, mutagenic homing expands the repertoire of potential host ligands used for attachment, a massive selective advantage [1, 49].

Since their initial discovery in *Bordetella bronchiseptica*, DGRs have been identified in more than 20 phyla of bacteria and archaea. These elements are found in a diverse number of hosts which occupy the gamut of environments and lifestyles, demonstrating their wide dispersal and suggesting they are of general utility. We identified DGRs in at least two *Legionella* species and have begun to characterize these elements. By analyzing the genetic components required for mutagenic homing we propose that the *Legionella*, and likely all bacterial DGRs function using the same conserved mechanism as the BPP phage, a process currently proposed to require template primed reverse-transcription [2]. We identified that levels of mutagenic homing are partially modulated by *Lp* global regulatory circuits which are essential for phenotypic transitions in response to stress [11] making the diversification of target genes in response to one or more specific stresses a likely scenario. Analysis of the *Lp* Corby target protein gene, *ldtA*, revealed a non-canonical bipartite N-terminal signal peptide that is responsible for its localization upon the bacterial surface. While additional experiments are required, it appears surface display requires the TAT system and likely lipoprotein sorting systems [21]. These pathways are widely conserved

suggesting surface display of lipoproteins may be a common property of bacteria [27]. DGRs are distributed across multiple strains of *Lp* and a strain of *Lt* share common features. They appear to reside within the same horizontally acquired genomic island and have nearly identical DGR diversification machinery. Despite these similarities, *Legionella* DGRs diversify a small family of TPs and the VR of each strain contains a unique pattern of adenine mutagenesis. These data suggests these elements are being maintained and that diversification of their TP is an individualized response to their environments that have been selected for. Furthermore, this work suggests that surface display of target proteins may be a common theme among bacterial DGRs. We have evaluated the hypothesis that *Lp* DGRs are active elements that have been exploited for surface display of variable proteins and a parallel between DGR activity and diversification of immunoglobulin scaffolds during mammalian immune responses can be drawn. Both require genetic mechanisms responsible for creating diversity which have co-evolved with protein scaffolds that display it, the immunoglobulin fold for antibodies and T-cell receptors and the CLec fold for DGRs. Both systems follow a primer of gene diversification, display of variable proteins followed by selection and amplification. However, unlike B cells, the Corby DGR is likely constrained by the fact that unmitigated diversification is deleterious and mutagenic homing is a response that is utilized at a precise time and place to maximize the benefit conferred by accelerated evolution.

References

1. Liu, M., et al., *Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.* Science, 2002. **295**(5562): p. 2091-4.
2. Guo, H., et al., *Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification.* Mol Cell, 2008. **31**(6): p. 813-23.
3. Miller, J.L., et al., *Selective ligand recognition by a diversity-generating retroelement variable protein.* PLoS Biol, 2008. **6**(6): p. e131.
4. Schillinger, T., et al., *Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF.* BMC Genomics, 2012. **13**: p. 430.
5. Minot, S., et al., *Hypervariable loci in the human gut virome.* Proc Natl Acad Sci U S A, 2012. **109**(10): p. 3962-6.
6. Medhekar, B. and J.F. Miller, *Diversity-generating retroelements.* Curr Opin Microbiol, 2007. **10**(4): p. 388-95.
7. Sidhu, S.S., et al., *Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions.* J Mol Biol, 2004. **338**(2): p. 299-310.
8. Alder, M.N., et al., *Diversity and function of adaptive immune receptors in a jawless vertebrate.* Science, 2005. **310**(5756): p. 1970-3.
9. Hairul Bahara, N.H., et al., *Phage display antibodies for diagnostic applications.* Biologicals, 2013. **41**(4): p. 209-16.
10. Loset, G.A. and I. Sandlie, *Next generation phage display by use of pVII and pIX as display scaffolds.* Methods, 2012. **58**(1): p. 40-6.
11. Dalebroux, Z.D., et al., *ppGpp conjures bacterial virulence.* Microbiol Mol Biol Rev, 2010. **74**(2): p. 171-99.
12. Dalebroux, Z.D. and M.S. Swanson, *ppGpp: magic beyond RNA polymerase.* Nat Rev Microbiol, 2012. **10**(3): p. 203-12.
13. Kuroda, A., et al., *Role of inorganic polyphosphate in promoting ribosomal protein degradation by the Lon protease in E. coli.* Science, 2001. **293**(5530): p. 705-8.
14. Jules, M. and C. Buchrieser, *Legionella pneumophila adaptation to intracellular life and the host response: clues from genomics and transcriptomics.* FEBS Lett, 2007. **581**(15): p. 2829-38.
15. Rasis, M. and G. Segal, *The LetA-RsmYZ-CsrA regulatory cascade, together with RpoS and PmrA, post-transcriptionally regulates stationary phase activation of Legionella pneumophila Icm/Dot effectors.* Mol Microbiol, 2009. **72**(4): p. 995-1010.
16. Wang, J.D., G.M. Sanders, and A.D. Grossman, *Nutritional control of elongation of DNA replication by (p)ppGpp.* Cell, 2007. **128**(5): p. 865-75.
17. Cazalet, C., et al., *Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity.* Nat Genet, 2004. **36**(11): p. 1165-73.
18. Dalebroux, Z.D., et al., *Distinct roles of ppGpp and DksA in Legionella pneumophila differentiation.* Mol Microbiol, 2010. **76**(1): p. 200-19.

19. Gordo, I., L. Perfeito, and A. Sousa, *Fitness effects of mutations in bacteria.* J Mol Microbiol Biotechnol, 2011. **21**(1-2): p. 20-35.
20. Beckwith, J., *The Sec-dependent pathway.* Res Microbiol, 2013. **164**(6): p. 497-504.
21. Palmer, T. and B.C. Berks, *The twin-arginine translocation (Tat) protein export pathway.* Nat Rev Microbiol, 2012. **10**(7): p. 483-96.
22. Shruthi, H., M.M. Babu, and K. Sankaran, *TAT-pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes.* J Mol Evol, 2010. **70**(4): p. 359-70.
23. Gralnick, J.A., et al., *Extracellular respiration of dimethyl sulfoxide by Shewanella oneidensis strain MR-1.* Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4669-74.
24. Widdick, D.A., et al., *The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17927-32.
25. Gimenez, M.I., K. Dilks, and M. Pohlschroder, *Haloferax volcanii twin-arginine translocation substates include secreted soluble, C-terminally anchored and lipoproteins.* Mol Microbiol, 2007. **66**(6): p. 1597-606.
26. Arambula, D., et al., *Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement.* Proc Natl Acad Sci U S A, 2013. **110**(20): p. 8212-7.
27. Okuda, S. and H. Tokuda, *Lipoprotein sorting in bacteria.* Annu Rev Microbiol, 2011. **65**: p. 239-59.
28. Linder, M.E. and R.J. Deschenes, *Palmitoylation: policing protein stability and traffic.* Nat Rev Mol Cell Biol, 2007. **8**(1): p. 74-84.
29. Chatzi, K.E., et al., *Breaking on through to the other side: protein export through the bacterial Sec system.* Biochem J, 2013. **449**(1): p. 25-37.
30. Paetzel, M., et al., *Signal peptidases.* Chem Rev, 2002. **102**(12): p. 4549-80.
31. Ursell, T.S., et al., *Analysis of surface protein expression reveals the growth pattern of the gram-negative outer membrane.* PLoS Comput Biol, 2012. **8**(9): p. e1002680.
32. Shruthi, H., et al., *Twin arginine translocase pathway and fast-folding lipoprotein biosynthesis in E. coli: interesting implications and applications.* Mol Biosyst, 2010. **6**(6): p. 999-1007.
33. Morales, V.M., A. Backman, and M. Bagdasarian, *A series of wide-host-range low-copy-number vectors that allow direct screening for recombinants.* Gene, 1991. **97**(1): p. 39-47.
34. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Mol Syst Biol, 2006. **2**: p. 2006 0008.
35. Maksimow, M., et al., *Simultaneous detection of bacteria expressing GFP and DsRed genes with a flow cytometer.* Cytometry, 2002. **47**(4): p. 243-7.
36. Guzman, L.M., et al., *Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter.* J Bacteriol, 1995. **177**(14): p. 4121-30.
37. Edwards, R.A., L.H. Keller, and D.M. Schifferli, *Improved allelic exchange vectors and their use to analyze 987P fimbria gene expression.* Gene, 1998. **207**(2): p. 149-57.

38. Tanaka, K., S.I. Matsuyama, and H. Tokuda, *Deletion of lolB, encoding an outer membrane lipoprotein, is lethal for Escherichia coli and causes accumulation of lipoprotein localization intermediates in the periplasm.* J Bacteriol, 2001. **183**(22): p. 6538-42.
39. Cowles, C.E., et al., *The free and bound forms of Lpp occupy distinct subcellular locations in Escherichia coli.* Mol Microbiol, 2011. **79**(5): p. 1168-81.
40. Aschtgen, M.S., et al., *SciN is an outer membrane lipoprotein required for type VI secretion in enteroaggregative Escherichia coli.* J Bacteriol, 2008. **190**(22): p. 7523-31.
41. Tseng, T.T., B.M. Tyler, and J.C. Setubal, *Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology.* BMC Microbiol, 2009. **9 Suppl 1**: p. S2.
42. Pearce, M.M. and N.P. Cianciotto, *Legionella pneumophila secretes an endoglucanase that belongs to the family-5 of glycosyl hydrolases and is dependent upon type II secretion.* FEMS Microbiol Lett, 2009. **300**(2): p. 256-64.
43. Korotkov, K.V., M. Sandkvist, and W.G. Hol, *The type II secretion system: biogenesis, molecular architecture and mechanism.* Nat Rev Microbiol, 2012. **10**(5): p. 336-51.
44. Yu, V.L., et al., *Distribution of Legionella species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey.* J Infect Dis, 2002. **186**(1): p. 127-8.
45. Pagnier, I., et al., *Genome sequence of Legionella tunisiensis strain LegM(T), a new Legionella species isolated from hypersaline lake water.* J Bacteriol, 2012. **194**(21): p. 5978.
46. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2010. **38**(Database issue): p. D5-16.
47. Kozak, N.A., et al., *Distribution of lag-1 alleles and sequence-based types among Legionella pneumophila serogroup 1 clinical and environmental isolates in the United States.* J Clin Microbiol, 2009. **47**(8): p. 2525-35.
48. Gogvadze, E. and A. Buzdin, *Retroelements and their impact on genome evolution and functioning.* Cell Mol Life Sci, 2009. **66**(23): p. 3727-42.
49. Doulatov, S., et al., *Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements.* Nature, 2004. **431**(7007): p. 476-81.