Mobile Genetic Elements 2:6, 287–291; November/December 2012; © 2012 Landes Bioscience

The low incidence of diversity-generating retroelements in sequenced genomes

Thomas Schillinger and Nora Zingler*
Department of Molecular Genetics; University of Kaiserslautern; Kaiserslautern, Germany

The insertion of a retrotransposable element is usually associated with adverse or, at best, neutral effects on the host. Diversity-generating retroelements (DGRs) are the first elements that seem to offer a direct selective advantage to their phage or prokaryote host by exact replacement of a short, defined region of a host gene with a hypermutated variant. In a previous study, we presented the software DiGReF for identification of DGRs in genome sequences, and compiled the first comprehensive set of diversity-generating retroelements in public databases. We identified 155 elements in more than 6,000 prokaryotic and phage genomes, which was a surprisingly low number. In this commentary, we will discuss the low incidence of these elements and speculate about the biological role of bacterial DGRs.

Keywords: diversity-generating retroelements, DiGReF, phages, bioinformatics, formylglycine-generating enzyme, database bias

Submitted: 10/30/12

Revised: 12/12/12

Accepted: 12/12/12

http://dx.doi.org/10.4161/mge.23244

*Correspondence to: Nora Zingler;
Email: nora.zingler@biologie.uni-kl.de

Commentary to: Schillinger T, Lisfi M, Chi J, Cullum J, Zingler N. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. BMC Genomics 2012; 13:430; PMID:22928525; http://dx.doi.org/10.1186/1471-2164-13-430

Classical retroelements spread throughout genomes by a copy-and-paste mechanism. The element is first transcribed into an RNA, which is then reverse transcribed by an element-encoded reverse transcriptase. The resulting cDNA duplicate is inserted in the host genome. If insertion occurs in non-coding and non-regulatory loci of the genome, there is usually no impact on the host phenotype. In contrast, an insertion in protein-encoding or regulatory regions often has drastic effects on the expression, folding, and function of proteins. As these changes are generally extremely deleterious to the host, retroelement activity is usually tightly controlled in somatic cells. Recent publications associate the activity of these elements with the pathogenesis of common diseases such as cancer and obesity.[1,2]

In contrast, a novel type of retroelement, termed a diversity-generating retroelement (DGR), employs a copy-and-replace mechanism.[3] First, an RNA is transcribed from the template repeat (TR). Using this RNA as a template, a reverse transcriptase introduces point mutations exclusively at adenine positions while synthesizing the cDNA. This cDNA is very similar to a region of the target host gene which is physically close to the DGR (**Fig. 1**). The mutagenized cDNA replaces the corresponding target DNA strand in a reaction that has been termed mutagenic homing.[4] In the prototypical DGR of *Bordetella* bacteriophage BPP1, described in pioneering publications by Jeff F. Miller and colleagues, the hypermutated host gene encodes a protein that is involved in attachment to target cells and therefore a key component for successful infection.[3-5] Protein variants with altered biochemical properties enable interaction with (and thus attachment to) a more diverse set of target cell surface proteins; this effectively broadens the host spectrum and thus confers an immediate advantage to the phage. The maximum number of possible variants correlates with the number of adenine residues in the hypermutated region; in the case of the *Bordetella* bacteriophage BPP1, a total of 23 adenines give rise to 4^{23} (or ~10^{14}) different DNA sequence variants, which is on par with the receptor variety of the human immune system.[3,6] Therefore, DGRs may allow the host to adapt to new or changing conditions through accelerated and targeted microevolution.

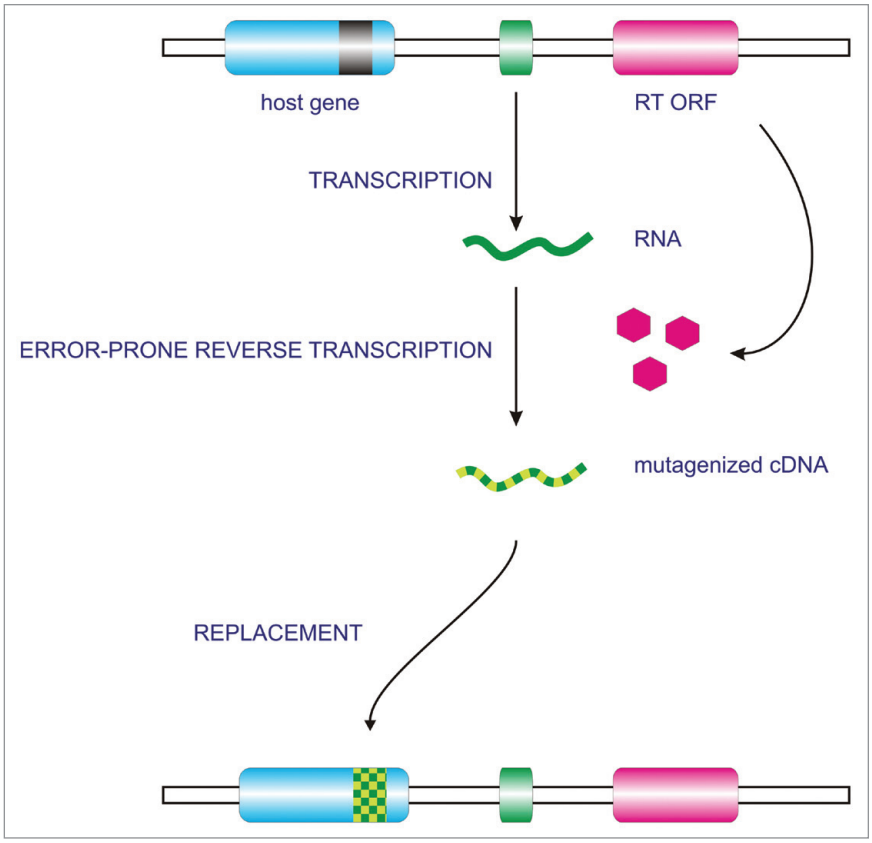

Figure 1. Mode of action of a diversity-generating retroelement. The prototypical DGR features an ORF encoding a reverse transcriptase (depicted in pink), a host gene (blue) with a variable region (black) and a template region (green). Transcription of an RNA from the template repeat is followed by reverse transcription in which mutations at adenine residues are introduced. The resulting cDNA replaces the variable region in the host gene, thereby mutagenizing the encoded protein. The process can be repeated for an unlimited number of rounds, as the template for transcription is maintained.

Diversity-Generating Retroelements Show a Low Incidence Among Sequenced Genomes

In the seminal paper of Doulatov et al.,[4] which presented a first overview of the distribution of DGRs among prokaryotes eight years ago, DGRs were found in a wide array of prokaryotes as diverse as cyanobacteria, green sulfur bacteria and gut bacteria. In our recent study, we expanded these findings by automating the search for DGRs. This now allows us to easily process the growing body of sequencing data supplied by next-generation sequencing techniques. In the present study, we screened an NCBI database that includes ~6,000 sequenced microbial genomes and ~600 dsDNA phage genomes. We identified DGRs in 152 bacteria species and 3 phages, while none were found in archaea or in eukaryotes.[7] Given the potentially large selective advantage of DGRs, we did not expect that less than 3% of sequenced prokaryotes and phages would contain them. As we have described in Schillinger et al.,[7] we utilized a very broad search approach including unusual reverse transcriptases and non-canonical mutation patterns, but we did not find any additional DGR-like elements. Therefore, we are quite confident that we did not overlook a major subpopulation of DGRs in the database. Here we discuss three possible reasons for the low incidence of DGRs. A combination of these explanations is also conceivable as they are not mutually exclusive.

Limited use of DGRs. The number of identified DGRs may only seem low at first glance. Even CRISPRs, which play an important role in prokaryotic immunity to phages or plasmids and have a much more versatile function than DGRs, are by no means ubiquitous in bacteria (prevalence of ~40%).[8] DGRs must meet several requirements that may be difficult to fulfill simultaneously in most organisms: (1) associate with a host gene in need of constant adaptation; (2) cause mutations that only diversify the target protein and do not completely disrupt it; and (3) evade host defense mechanisms when coming from an exogenous source.

Aside from these tight constraints on the expansion of DGRs, they may simply be a young class of mobile elements that has not had sufficient time to spread. However, previous studies and our own results (see below) demonstrate that target proteins of DGRs display a very high diversity. If DGR acquisition were prohibitively complicated or very recent, why do we not observe more homogenous groups of target proteins, descending from only a few ancestor proteins?

Database bias. For our previous analysis, we used the non-redundant (nr) protein database from NCBI, which essentially includes deposited protein sequences and translated coding sequences of all deposited genomes with the exception of environmental metagenomic data. For historical and practical reasons, sequences in the nr database often come from organisms that are medically or industrially relevant and easy to cultivate; they do not necessarily represent a properly randomized sample of all existing organisms. More importantly, most fully sequenced prokaryotes have been maintained in culture for up to several decades under optimal growth conditions and without selective pressure. Extended maintenance under laboratory conditions eliminates the need for mechanisms to support fast adaptation to changing environmental conditions, and has often been shown to lead to rapid loss of virulence.[9,10] If DGRs should be important for cell-to-cell communications in bacterial consortia, monoculture would also render this function superfluous. Many DGRs in cultured strains may therefore be lost after several generations or too divergent to be recognized by our program.

To assess whether DGRs are more abundant in free-living bacteria, we performed a PHI-BLAST search[11] for RTs with a polypeptide motif characteristic for DGRs ([LIV]GxxxSQ) in the metagenomic protein database (env_nr). This database at present contains approximately six million translations from environmental DNA sequence data. We found 106 additional candidate DGR-RTs that are distinct from the nr-set (data not shown). However, it was impossible to confirm that these RTs belong to complete DGRs because the sequence fragments in env_nr are usually too short. Also, a direct comparison of the number of DGR hits found in the nr and the env_nr database is not feasible due to the multiple differences in composition and size of the databases. Still, finding >100 additional potential DGRs so easily in these low quality metagenomes strongly suggests a higher prevalence of DGRs in free-living organisms.

Phage association. Another possibility we explored was the theory that DGRs only occur in phages, and DGRs in prokaryotic genomes are components of prophages instead of integral parts of the host genomes. Indeed, five prokaryotic target proteins that we found had been annotated with "major tropism determinant," and manual examination of their genomic vicinity revealed additional phage proteins, suggesting the presence of prophages in the host genomes. In further analyses, we used the ACLAME database,[12] which provides pre-mapped prophages in genomes of sequenced organisms, including 13 organisms with DiGReF-confirmed DGRs. In nine of these, the DGR element was part of a tagged prophage region. Using ACLAME's Prophinder tool, 68 of the 161 prokaryotic potential target proteins yield matches with prophage-related proteins (E-value cutoff 0.0001). However, prophage prediction tools may make some errors. A list of 161 random proteins from the same hosts also returned a considerable number of hits (21.7% vs. 42.2%), challenging the reliability and significance of the results. Still, the true number of phage-associated DGRs, and thus their fraction in phages, may be significantly higher than current investigations indicate. An **exclusive** phage association however does not seem likely, given our observations that DGRs can also be found on plasmids, transposons or other non-phage mobile genetic elements, and that transfer of DGRs between prokaryotes seems to take places using these vectors.[7] This promiscuous hitchhiking behavior of DGRs fits in well with the modular nature of most mobile genetic elements that share and exchange various components of phages, plasmids and transposons.[13-15] DGRs might have a place in the mosaic of mobile elements by acting as a "fitness module" that provides an advantage to any host, not just viruses.

Table 1. Annotations of potentially variable proteins

Annotation	Count	Fraction (%)
Hypothetical/predicted protein, Unknown function	82	50.0
FGE sulfatase enzyme	45	27.4
DUF1566	7	4.3
Major tropism determinant	5	3.0
Concanavalin A-like lectin/glucanases superfamily	5	3.0
DUF3988	2	1.2
Other	18	11.0

Target Proteins of Diversity-Generating Retroelements

To better understand the origin and prevalence of DGRs in bacteria, it would be useful to know the exact purpose they serve in their respective host organisms, or, more essentially, the function of the mutated target genes. Previous studies have proposed prey and predator characteristics for proteins that are hypermutated by DGRs, which means that proteins either have to bind diverse targets (predators) or are the respective binding targets themselves (prey), and must diversify to escape binding by predators.[16] We therefore used our data set for retrieval of the sequence and the corresponding annotations of 164 potential DGR target proteins from NCBI Protein database. Next, we evaluated the obtained data under aspects of function and phylogenetic relations.

As most of the genomes that have been deposited in the past few years are poorly annotated, 55.5% of the retrieved target proteins are merely tagged as "hypothetical," "predicted" or "domain of unknown function (DUF)" (**Table 1**), providing no immediate clue to their function. Another large group (~27%) has been associated with the formylglycine sulfatase superfamily (NCBI Conserved Domain Database CDD cl15394), a class of eukaryotic proteins that generates formylglycine at the active site in sulfatases. 3% of all target proteins are homologs of the major tropism determinant of *Bordetella* phage, and the remainder are annotated with various descriptions such as "DNA topoisomerase IV" or "serralysin," which do not hold up upon closer inspection.

A simple phylogenetic tree based on a ClustalW alignment of the target proteins shows three major branches (**Fig. 2**). All proteins of the "FGE sulfatase enzyme" group can be found in the same branch as "major tropism determinant" proteins and "Concanavalin A-like lectin/glucanases" superfamily proteins, indicating a certain level of relatedness of these proteins, while the other two branches are largely composed of proteins without annotation or likely wrong annotations. The level of conservation in these branches is very low (BLASTp E-values often worse than 10^{-2}), suggesting that DGRs have acquired a wide array of different target molecules, consistent with previous observations.[16]

The lack of informative annotations highlights a general caveat of the massive increase in sequence data that we have witnessed in the past decade. Although there are breathtaking advances in sequencing technology, processing of sequencing data and the automated annotation process are bottlenecks which hinder the full utilization of the accumulated data.[17,18] Misannotations occur through comparing newly sequenced genomes and translated open reading frames to their closest relatives. However, a relationship does not necessarily imply homologous function, especially when the conservation is low and a protein is annotated

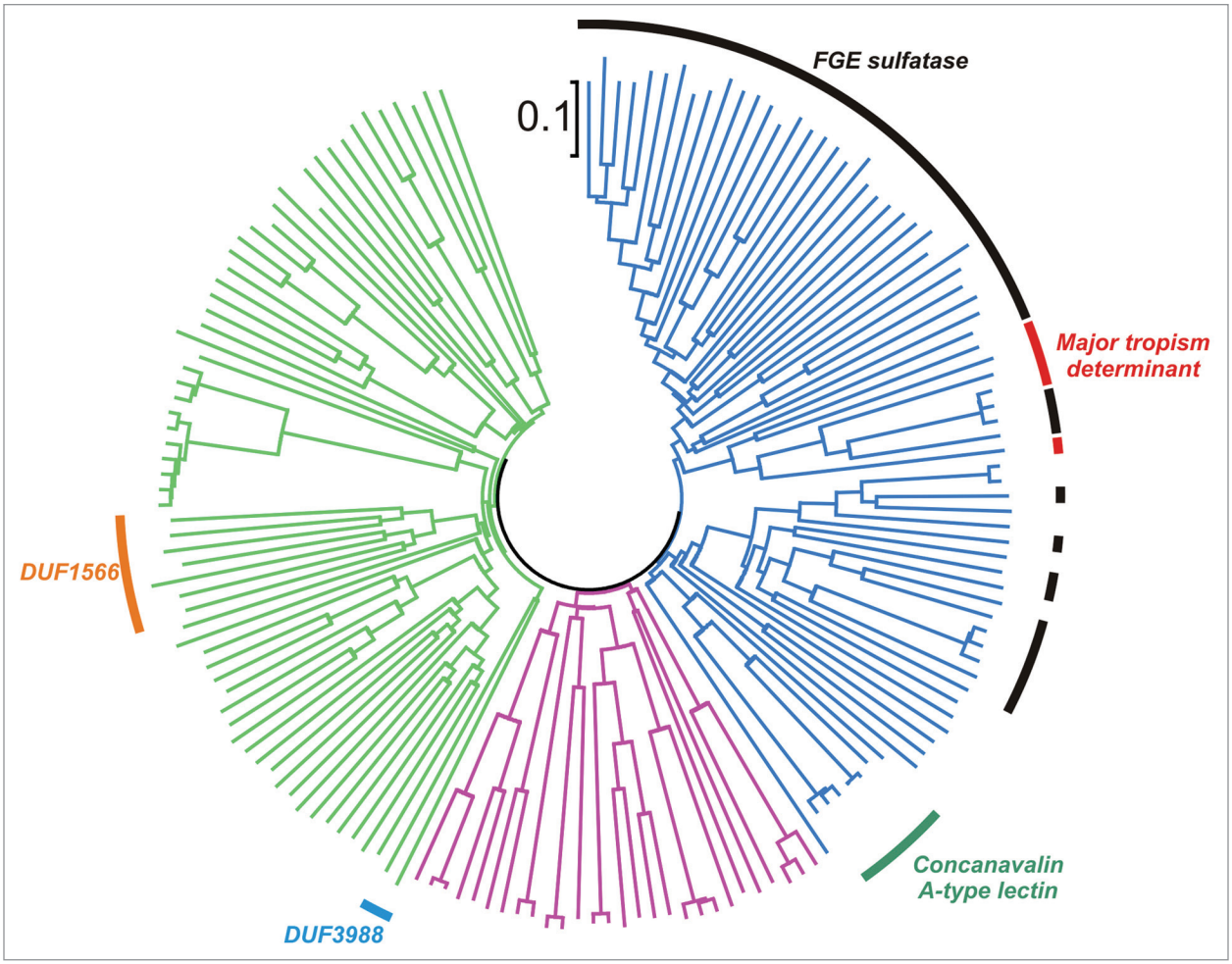

Figure 2. ClustalW tree of 164 DGR target genes. Protein sequences of DGR target genes were extracted and aligned using the ClustalW algorithm. The scale bar indicates amino acid substitutions per site as a measure of distance. The tree is divided into three branches shown in different colors. Arcs of different colors cover all entries that share a common annotation (see **Table 1**). The blue branch of the tree consists mostly of proteins with FGE sulfatase annotation, while also including several proteins tagged as "major tropism determinant" and "concanavalin A-type lectin," suggesting a common lectin-type fold of these proteins. Proteins in the pink and green branches are predominantly annotated as proteins of unknown function.

from comparison to a similar protein just because no real or better homolog has been found and described yet.

The DGR target proteins associated with the formylglycine sulfatase superfamily may be a perfect example for this. Almost all relevant research on this class of proteins was done in eukaryotes. While sulfatases and sulfatase-modifying enzymes have been identified in prokaryotes as well,[19] it is doubtful that a catalytic enzyme of such a specific function needs to be diversified by DGRs. Le Coq and Ghosh proposed a common ancestor protein of genuine FGE proteins and DGR proteins like Mtd from *Bordetella* bacteriophage or TvpA from *Treponema denticola*.[20] It is easily conceivable that some DGR target proteins share some structural features with FGE proteins such that they can both recognize specific oligopeptide motifs. Assuming a protein recognition rather than an oxidizing function makes more sense in the DGR context, and in TvpA, one of the three catalytic residues is actually mutated and catalytic activity abolished by DGR activity.[20] Considering that a number of FGE-like target proteins display additional domains such as serine/threonine kinases or NACHT NTPase, we can even speculate that these proteins play a role in signal transduction. Prokaryotic serine/threonine kinases have been associated with a wide array of cellular mechanisms, including stress response and secretion of virulence factors, and it seems conceivable that participation of a diversity-generating retroelement supports adaptation to environmental stimuli. However, without the assistance of classical wet lab experiments comprising biochemistry, physiology and genetics, such considerations remain speculation. Data mining is a powerful tool that allows fascinating insights into biological correlations, but it reaches its limits when dealing with entirely novel systems such as DGRs.

Conclusion

In this commentary, we examined the surprisingly low incidence of DGRs in sequenced genomes, despite their potential to confer selective advantages to the hosts. Currently this discussion remains highly speculative, as we lack suitable sequencing data from environmental samples

and refined tools for prophage prediction. Likewise, the purpose of identified DGRs remains elusive, as the function of their target proteins is still unknown. This dearth of data emphasizes that although high-throughput sequencing and bioinformatics have provided us with unprecedented analytical possibilities, they need to be grounded on a solid foundation of microbiological and biochemical research to minimize the current database bias for culturable microorganisms and the annotation bias toward well-characterized protein families. This is true not only for DGRs, but for processing the vast amounts of sequencing data in general. The next years will surely bring a new level of data networking in biological sciences, connecting results from next generation sequencing projects with wet lab experiments. Surely such advances will provide more satisfying insights into DGR evolution and function which will enable us to harness their fascinating novel features for drug development, phage therapy, and countering the rapid evolution of antibiotic resistance.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We would like to thank A. Solem for critical reading of the manuscript. This work has been supported by a grant from the EU-FP7 program (Marie Curie International Reintegration grant PIRG05-GA-2009–248023) to N.Z.

References

1. Kuehnen P, Mischke M, Wiegand S, Sers C, Horsthemke B, Lau S, et al. An Alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity. PLoS Genet 2012; 8:e1002543; PMID:22438814; http://dx.doi.org/10.1371/journal.pgen.1002543.
2. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, et al.; Cancer Genome Atlas Research Network. Landscape of somatic retrotransposition in human cancers. Science 2012; 337:967-71; PMID:22745252; http://dx.doi.org/10.1126/science.1222077.
3. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, et al. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science 2002; 295:2091-4; PMID:11896279; http://dx.doi.org/10.1126/science.1067467.
4. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, et al. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 2004; 431:476-81; PMID:15386016; http://dx.doi.org/10.1038/nature02833.
5. Guo H, Tse LV, Barbalat R, Sivaamnuaiphorn S, Xu M, Doulatov S, et al. Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. Mol Cell 2008; 31:813-23; PMID:18922465; http://dx.doi.org/10.1016/j.molcel.2008.07.022.
6. Medhekar B, Miller JF. Diversity-generating retroelements. Curr Opin Microbiol 2007; 10:388-95; PMID:17703991; http://dx.doi.org/10.1016/j.mib.2007.06.004.
7. Schillinger T, Lisfi M, Chi J, Cullum J, Zingler N. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. BMC Genomics 2012; 13:430; PMID:22928525; http://dx.doi.org/10.1186/1471-2164-13-430.
8. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 2009; 34:401-7; PMID:19646880; http://dx.doi.org/10.1016/j.tibs.2009.05.002.
9. Chosa H, Makino S, Sasakawa C, Okada N, Yamada M, Komatsu K, et al. Loss of virulence in Shigella strains preserved in culture collections due to molecular alteration of the invasion plasmid. Microb Pathog 1989; 6:337-42; PMID:2770505; http://dx.doi.org/10.1016/0882-4010(89)90075-2.
10. Domenech P, Reed MB. Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from Mycobacterium tuberculosis grown in vitro: implications for virulence studies. Microbiology 2009; 155:3532-43; PMID:19661177; http://dx.doi.org/10.1099/mic.0.029199-0.
11. Zhang Z, Schäffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 1998; 26:3986-90; PMID:9705509; http://dx.doi.org/10.1093/nar/26.17.3986.
12. Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res 2010; 38(Database issue):D57-61; PMID:19933762; http://dx.doi.org/10.1093/nar/gkp938.
13. Osborn AM, Böltner D. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid 2002; 48:202-12; PMID:12460536; http://dx.doi.org/10.1016/S0147-619X(02)00117-8.
14. Toussaint A, Merlin C. Mobile elements as a combination of functional modules. Plasmid 2002; 47:26-35; PMID:11798283; http://dx.doi.org/10.1006/plas.2001.1552.
15. Wozniak RA, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 2010; 8:552-63; PMID:20601965; http://dx.doi.org/10.1038/nrmicro2382.
16. McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, Marti-Renom MA, et al. The C-type lectin fold as an evolutionary solution for massive sequence variation. Nat Struct Mol Biol 2005; 12:886-92; PMID:16170324; http://dx.doi.org/10.1038/nsmb992.
17. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 2012; 13:667-72; PMID:22898652; http://dx.doi.org/10.1038/nrg3305.
18. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. Brief Bioinform 2012; PMID:22408191; http://dx.doi.org/10.1093/bib/bbs007.
19. Bakal CJ, Davies JE. No longer an exclusive club: eukaryotic signalling domains in bacteria. Trends Cell Biol 2000; 10:32-8; PMID:10603474; http://dx.doi.org/10.1016/S0962-8924(99)01681-5.
20. Le Coq J, Ghosh P. Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. Proc Natl Acad Sci U S A 2011; 108:14649-53; PMID:21873231; http://dx.doi.org/10.1073/pnas.1105613108.