The Journal of Microbiology (2011) Vol. 49, No. 1, pp. 35-45

DOI 10.1007/s12275-011-1024-2

Reductive Divergence of Enterobacterial Repetitive Intergenic Consensus Sequences among *Gammaproteobacteria* Genomes

Young-Gun Zo

Department of Biology, Kyungsung University, Busan 608-736, Republic of Korea

(Received January 13, 2011 / Accepted January 27, 2011)

Enterobacterial repetitive intergenic consensus (ERIC) sequence is a transcription-modulating, nonautonomous, miniature inverted-repeat transposable element. Its origin and the mechanism of highly varying incidences, limited to *Enterobacteriaceae* and *Vibrionaceae*, have not been identified. In this study, distribution and divergence of ERICs along bacterial taxonomic units were analyzed. ERICs were found among five families of gammaproteobacteria, with the copy numbers varying with exponential increments. The variability was explained by genus (45%) and species (36%) affiliations, indicating that copy numbers are specific to subfamily taxa. ERICs were interspersed in genomes with considerable divergences. Locations of ERICs in a genome appeared to be strongly conserved in a strain, moderately in a species or a genus, and weakly in a family. ERICs in different species of a genus were from the identical population of sequences while ERICs in different genera of a family were nearly identical. However, ERICs in different families formed distinct monophylectic groups, implying vertical transmission of diverging population of sequences. In spite of large difference in copy numbers, overall intra-genome evolutionary distances among ERICs were similar among different species, except for a few genomes. The exceptions substantiated hypotheses of genetic drifts and horizontal gene transfers of mobility capacity. Therefore, the confined, variable distribution of ERIC could be explained as a two-step evolution: introduction and proliferation of ERIC in one of the progenitors of gammaproteobacteria, followed by vertical transmission under negative selection. Deterioration of sequences and reduction in copy number were concluded to be the predominant patterns in the evolution of ERIC loci.

***Keywords*:** Class 2 transposon, intergenic sequences, interspersed repeated sequences, MITE, analysis of molecular variance

Interspersed repeated sequences (IRSs) are found in most genomes of eukaryotes and prokaryotes. As exemplified from the primate-specific *Alu*, distribution of a given family of IRS is often taxon-specific (Cordaux and Batzer, 2009). IRSs cause genome instability and function as a driver for emergence of new genes and genome organizations (Lewis *et al.*, 1999; Achaz *et al.*, 2003; Treangen *et al.*, 2009), hence providing substrates and devices for taxon-specific evolution of genomes, i.e., speciation. Their contributions to genome plasticity and evolution were evident in comparative genomics surveys (Achaz *et al.*, 2003; Delihas, 2008; Treangen *et al.*, 2009).

Dozens of IRS families were identified in prokaryotic genomes (Versalovic and Lupski, 1998; Bachellier *et al.*, 1999; Delihas, 2008; Treangen *et al.*, 2009), and more are expected to be discovered as more genomes are sequenced. While cellular functions and evolutionary impacts are not well characterized for most of IRS families, some of the mobile forms, including repetitive extragenic palindromic (REP) elements, bacterial interspersed mosaic elements (BIME), *Rickettsia* palindromic elements (RPEs) and enterobacterial repetitive intergenic consensus sequences (ERICs), were better characterized (Delihas, 2008).

REP elements are two inverted sequences separated by a integration host factor recognition sequence (Oppenheim *et al.*, 1993) and DNA targets for Insertion Sequences (Tobes and Pareja, 2006). REP elements are transposition hot spots conforming genomic plasticity via mobile elements and often arranged in composite repetitive structures, i.e., BIMEs. BIMEs are known to stabilize mRNA of upstream ORF (Bachellier *et al.*, 1999). More importantly, they provoke pause in RNA polymerase movement and Rho-dependent transcription attenuation. REP element and BIME origins are suggested to be a transposon mobilized by IS200/IS605 tyrosine transposases, which was found to be physically and evolutionarily associated with REPs (Nunvar *et al.*, 2010). Therefore, REP and BIME are nonautonomous transposable elements (TEs) that have functions as DNA and mRNA. Being large number of them interspersed throughout a genome, they can have genome-wide effects on regulation of gene expression. RPEs were found only among the obligate intracellular parasite, *Rickettsia* spp. Unlike REPs, RPEs are inserted in-frame within functional open reading frames of *Rickettsia* encoding specific sequence of amino acids in the middle of many unrelated genes (Ogata *et al.*, 2000). Phylogenetic analysis revealed that RPE was introduced into the genus, subsequent to divergence of the genus from the other alphaproteobacteria (Amiri *et al.*, 2002).

ERIC sequences are 127 bp elements found in some species of *Enterobacteriaceae* and *Vibrionaceae* families of gammaproteobacteria (Versalovic and Lupski, 1998; Bachellier *et al.*,

* For correspondence. E-mail: zoyful@gmail.com; Tel: +82-51-663-4643; Fax: +82-51-627-4645

1999). Based on structural similarity of terminal inverted repeats which conforms a Class 2 TE, ERIC was characterized as a small nonautonomous mobile element, miniature inverted repeat transposable element (MITE) (De Gregorio *et al.*, 2005). Its function as genomic DNA is not known, but functions as RNA were well characterized empirically, together with *in silico*-characterized functions of multiple short ORFs (Delihas, 2007). ERICs are cotranscribed with upstream or downstream ORFs and stabilize/destabilize the transcript depending on its orientation and hairpin structure. It also has an RNase E cleavage site which is exposed during translation by location of ribosome (De Gregorio *et al.*, 2005). While most ERICs

Table 1. Numbers (mean±SD) of complete genome sequences (*N*), ERIC loci per genome (*L*), and ERIC loci per 1,000 genes (*D*)

Family	Species name	*N*	*L*	*D*	Accession no.
Enterobacteriaceae	*Citrobacter koseri*	1	9	1.8	CP000822
	Citrobacter rodentium	1	1	0.2	FN543502
	Cronobacter sakazakii	1	1	0.2	CP000783
	Cronobacter turicensis	1	4	0.9	FN543093
	Dickeya dadantii	3	15±9	3.4±2.0	CP002038, CP001836, CP001654
	Dickeya zeae	1	45	10.3	CP001655
	Enterobacter cloacae	3	2±1	0.5±0.2	CP002272, CP001918, FP929040
	Enterobacter sp. 638	1	8	1.9	CP000653
	Erwinia amylovora	2	4	1.1	FN666575, FN434113
	Erwinia billingiae	1	3	0.6	FP236843
	Erwinia pyrifoliae	2	5	1.3±0.1	FN392235, FP236842
	Erwinia sp. Ejp617	1	7	1.9	CP002124
	Escherichia coli	40	7±3	1.5±0.7	footenote[b]
	Escherichia fergusonii	1	9	2.0	CU928158
	Klebsiella pneumoniae	3	6±1	1.1±0.1	CP000964, AP006725, CP000647
	Klebsiella variicola	1	6	1.2	CP001891
	Pantoea vagans	1	1	0.3	CP002206
	Pectobacterium atrosepticum	1	76	16.8	BX950851
	Pectobacterium carotovorum	1	43	9.8	CP001657
	Pectobacterium wasabiae	1	37	7.8	CP001790
	Photorhabdus asymbiotica	1	366	82.6	FM162591
	Photorhabdus luminescens	1	145	25.3	BX470251
	Proteus mirabilis	1	4	1.1	AM942759
	Salmonella enterica	19	14±2	3.0±0.5	footenote[b]
	Serratia proteamaculans	1	11	2.2	CP000826
	Shigella boydii	2	8±1	1.6	CP001063, CP000036
	Shigella dysenteriae	1	3	0.6	CP000034
	Shigella flexneri	4	6	1.3	footenote[b]
	Shigella sonnei	1	6	1.3	CP000038
	Xenorhabdus nematophila	1	12	2.6	FN667742
	Yersinia enterocolitica	1	182	43.7	AM286415
	Yersinia pestis	10	59±3	14.7±1.5	footenote[b]
	Yersinia pseudotuberculosis	4	75±4	17.7±1.0	footenote[b]
Vibrionaceae[a]	*Vibrio cholerae*	5	101±2	25.4±1.0	footenote[b]
	Vibrio vulnificus	2	16±3	3.6±0.3	AE016795+6, BA000037+8
	Vibrio harveyi	1	1	0.2	CP000789+90
	Vibrio splendidus	1	2	0.4	FM954972+3
Aeromonadaceae	*Tolumonas auensis*	1	5	1.5	CP001616
Psychromonadaceae	*Psychromonas ingrahamii*	1	1	0.3	CP000510
Shewanellaceae	*Shewanella baltica*	2	1	0.2	CP000891, CP001252

[a] A *Vibrionaceae* bacterium has two chromosomes in its genome. Accession numbers for each pair of chromosomes were indicated as a pair. The last one or two digits in the accession numbers of the smaller chromosomes were different from those of the larger chromosomes in a genome. Accession numbers for the former were shown after a plus sign after the latter.

[b] Accession numbers: *E. coli*, FN554766, CP000247, AP010960, AP010953, AP010958, CU928145, CP001671, CP000468, CP000946, AM946981, CP001509, CP001665, CP001846, AE014075, CP001637, FM180568, CP000800, CP001164, CU928162, AE005174, FN649414, CP000802, CU928160, CU928164, CP001969, AP009048, CP000948, CP001396, U00096, CU651637, CP000819, CU928161, BA000007, AP009240, AP009378, CP000970, CP001368, CP002167, CU928163, CP000243, and CU928158; *S. enterica*, CP000880, CP001138, AE017220, CP001144, AM933172, AM933173, CP001120, CP001113, FM200053, CP000026, CP000886, CP000857, CP001127, AL513382, AE014613, CP001363, FN424405, AE006468, and FQ312003; *Y. pestis*, AE017042, CP000901, CP000308, AL590842, CP001585, CP001589, AE009952, CP000305, CP000668, and CP001593; *Y. pseudotuberculosis*, CP000720, BX936398, CP001048, and CP000950; *S. flexneri*, AE005674, CP000266, CP001383, and AE014073; *V. cholerae*, CP001233+4, CP001485+6, AE003852+3, CP000626+7, and CP001235+6.

were found at intergenic locations, some were intragenic with in-frame or off-frame with the fused ORF. The frame-shifted translation was predicted to provide the fused protein with a putative transmembrane domain (Delihas, 2007, 2008).

However, the multi-family distribution of ERIC among bacterial taxa raises intriguing questions about its origin and dynamics in gammaproteobacteria. While many *Enterobacteriaceae* species have dozens or hundreds of copies of ERIC, majority of genera in *Enterobacteriaceae* and *Vibrionaceae* have only a few ERICs or are lacking an ERIC. Interestingly, ERICs are abundant in *Vibrio cholerae* (about 100 per genome) while other *Vibrio* spp. are lacking or have only a few copies of ERIC. Because the scattered distribution of interspersed repeats contributes to structural dynamics of bacterial genome (Achaz *et al.*, 2003; Treangen *et al.*, 2009; Muñoz-López and García-Pérez, 2010), the origin and dynamics of IRSs should parallel the evolution of bacterial genomes. The condition of highly uneven gradient in the copy numbers of ERIC among diverse gammaproteobacteria provides a unique opportunity for elucidating evolutionary tracks of interspersed repeats of bacterial genomes.

In this study, distribution and divergence of ERIC, both within a genome and along taxonomic hierarchies in the bacteria kingdom, were analyzed based on a comprehensive survey on complete genome sequences in up-to-date public repositories. With the aim of revealing mechanisms for introduction and skewed distributions of ERICs, phylogenetic and biometrical analyses were performed on copy number, sequence polymorphism and intra-chromosomal locations of ERIC sequences in completely sequenced bacterial genomes.

Materials and Methods

Sequence acquisition

With the aim of analyzing all available genomes harboring ERIC sequences, entire nucleotide entries in the GenBank nr/nt database were searched for full-length or nearly-complete ERIC sequences by the BLAST server (Altschul *et al.*, 1997) in National Center for Biotechnology Information (http://blast.ncbi.nlm.nih.gov, U.S. National Library of Medicine, Maryland, USA) on Nov. 1, 2010. Two query ERIC sequences were employed, which were determined as the most frequent sequence of the genomes of *V. cholerae* N16961 or *Photorhabdus asymbiotica* ATCC 43949. BLAST hits of the two queries were pooled for further analysis. In a preliminary analysis employing ERIC sequences in *Escherichia coli* K12 as query sequences, the two query-source strains showed the most number of BLAST hits in families *Vibrionaceae* and *Enterobacteriaceae*, respectively, with >100 hit frequencies. BLAST analyses were performed with search word-size and expect value set as 7 and <0.01, respectively, as a parameter optimization for nearly full-length hits. The pooled BLAST hits were purged in two steps: selection of hits with >100 bp alignment with query sequences and selection of hits from completely sequenced genomes. According to GenBank entries of hit sequences, the size of genomes, the number of ORFs and the location of replication origins of genomes were determined (Table 1). For calculation of number of ERIC loci per genome, average number of loci per gene, or locations of each locus relative to the replication origin, those values of a genome were required. Therefore, the sequence must be of a complete genome, and hence unfinished whole genome shotgun sequences were excluded for analyses employed in this study.

Analysis of nucleotide sequence divergence

Average number of nucleotide differences per site between any two ERIC sequences randomly chosen from a given haploid genome was used as the estimate of nucleotide diversity of ERICs in a given genome (Nei and Kumar, 2000). In this study, the nucleotide diversity estimate was interpreted as an index of the magnitude of intra-genome diversity of ERICs.

For the purpose of analyzing evolutionary depth of ERIC sequence evolution, contribution of a given difference of standings in taxonomic hierarchies (family, genus, species, and strain) to genetic divergence of ERICs was examined by analysis of molecular variance (AMOVA) (Excoffier *et al.*, 1992) implemented in the Arlequin for Windows version 3.5 (Excoffier and Lischer, 2010). Because the current AMOVA framework is limited to population structures with up to two tiers of nested variance components, the contribution and significance of each source of variance were tested separately in four different structures (Table 2). The residual variance was set to be intra-genome variance of ERICs for all structures. To test contribution of difference in family, difference in genome was nested under each family (the structure FAM in Table 2). Likewise, estimation of contribution by differences in genus, species and strain were set to be the upper tier under which difference of genomes was nested (the structures GEN, SPP, and STR in Table 2). Significances of fixation index and variance were tested by 20,000 times of permutation.

For reconstruction of evolutionary path of diversification of ERICs, maximum likelihood (ML) method was used, employing PhyML version 3.0 (Guindon *et al.*, 2010). Transition/transversion ratio, proportion of invariable sites and Gamma shape parameter were estimated by PhyML based on sample data and the nucleotide substitution model of Hasegawa-Kishino-Yano (Hasegawa *et al.*, 1985). Starting tree search with BIONJ tree, 2,000 bootstrap dataset were analyzed to obtain branch support values. For analysis of intra-genome divergence of ERICs, all ERIC sequences from a genome was analyzed by the ML method. For analysis of inter-genome divergence, the centroid sequence of a given genome was selected to represent ERICs in each genome. The centroid sequence was determined as the sequence producing the maximum value in the sum of base matching scores in all non-redundant pairwise comparisons.

Analysis of variance and spatial correlation

To address the question whether location of ERICs in a pair of genomes are conserved, spatial correlation of ERICs on the circular chromosomes of bacteria was examined by estimating circular correlation coefficient (Jammalamadaka and SenGupta, 2001). A bacterial chromosome was divided into 512 equal-length sections, like arcs on a circle. Then, density of ERIC loci for each section, i.e., arc, was calculated as the number of ERIC loci per section. Because the location of replication origin and the direction of DNA strands were different among chromosomes, it was required to find the optimum geometric alignment for circular correlation by rotating one of the chromosomes into two directions. A total of 1024 circular correlation coefficients were calculated for each pair of chromosomes. The estimate of spatial correlation between a pair of chromosomes was determined as the maximum value among the 1024 circular correlation coefficients.

For the purpose of estimating contributions of taxonomic hierarchies to variability of the number and the locations of ERIC loci in a genome, analysis of variance (ANOVA) was performed employing a mixed model framework. Because the hierarchy of order, family, genus, species and strain were random effects with a sequentially

nested structure, the independent variables of ANOVA model were set as the nested structure of five random variables. To summarize dependence of ERIC abundance on taxonomic hierarchy, the copy number was normalized in respect to the genome size by dividing the copy number of ERICs by the number of ORFs of a given genome and then log-transformed. For ANOVA of circular correlation coefficient, the arcsine transformation was required to meet the normality assumption of ANOVA. Shapiro-Wilk's test was used to test normality of native and transformed data. Likelihood ratio (LR) test was used to test the significance of each explanatory variable, by comparing the likelihood values of ANOVA models with versus without the nested level of interest. All statistical procedures were performed in the statistical programming environment R (R Development Core Team, 2005), with the significance criterion set as the type I error rate at <0.05.

Results

Variation in number of ERIC locus per genome

Depending on taxonomic affiliation of bacterium, the number of ERIC sequence loci varied greatly. First of all, ERIC sequence was strongly confined to a specific class in the domain Bacteria. Among all microorganisms that had at least one nucleotide sequence in the nr/nt database of GenBank, only four orders under the class *Gammaproteobacteria*, namely the orders *Aeromonadales*, *Alteromonadales*, *Enterobacteriales*, and *Vibrionales*, carried ERIC. Those orders comprised five families in Table 1. The number of ERIC loci in the genomes of five families ranged from zero to 366. Although the majority of *Enterobacteriaceae* bacteria harbored ERIC sequences, complete genome sequences of the genera *Edwardsiella*, *Buchnera*, *Sodalis*, and *Wigglesworthia* did not have the repetitive sequence. According to the survey results, non-*Enterobacteriaceae* bacteria rarely harbor ERIC sequences. As it is expected, *Vibrionaceae* was the exception to this remark, only because of *V. cholerae*. In the genus *Vibrio*, complete sequences of genomes of *Photobacterium* and *Grimontia* species did not produce BLAST hits matching the search criteria. *Aliivibrio salmonicida* LFI1238 carried one hit with 114 bp alignment length. However, the inverted repeat part, the key feature of a typical MITE, was different from that of ERIC by 12 bp. Therefore, the *Vibrio* genus was unique to have >100 loci per genome in the family because of *V. cholerae*. In conclusion, presence/absence of ERIC sequence in bacterial genome generally depended on taxonomic standing of a bacterium, including its order, family, genus, and species designations. On the contrary, the number of ERIC loci within a species appeared to vary less extensively, as it was implied by the small standard deviations (≤ 4) for most species in Table 1. An exception to this general pattern was *Dickeya dadantii*. The three strains (Ech596, Ech703, and 3937) of the species showed 24, 6 and 15 ERIC loci, respectively.

To summarize the dependence of ERIC incidences on taxonomic hierarchy, mixed-model ANOVAs were performed by fitting log-normalized values of number of ERIC loci per genome to different sets of nested random effect variables comprising the taxonomic hierarchies from order, family, genus,

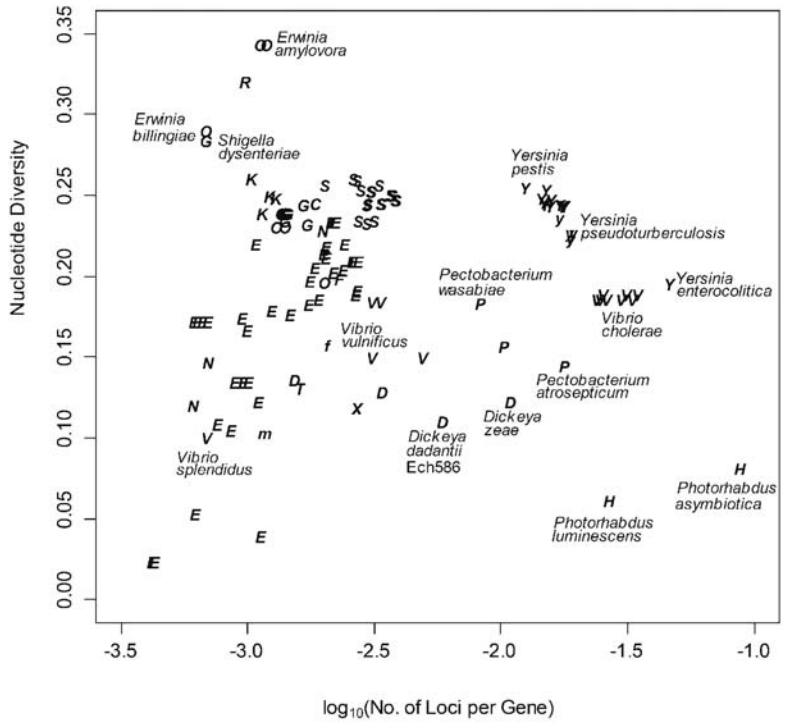

Fig. 1. Relationship between nucleotide diversity and number of ERIC loci per genome (Symbols for chromosomes of strains: *R*, *Cronobacter* spp.; *N*, *Enterobacter* spp.; *O*, *Erwinia* spp.; *A*, *Pantoea vagans* C9-1; *H*, *Photorhabdus* spp.; *i*, *Psychromonas ingrahamii* 37; *m*, *Proteus mirabilis* HI4320; *F*, *Serratia proteamaculans* 568; *b*, *Shewanella baltica*; *G*, *Shigella* spp.; *f*, *Escherichia fergusonii* ATCC 35469; *y*, *Yersinia pseudotuberculosis*; all other letters, the first letter of genus name. Species names were shown for distinctively located symbols).

species and strain. According to the result, the variability of ERIC loci incidences per genome did not depend on order or family affiliation of a genome (LR test; $P \geq 0.62$). The most proportion of the variability (45% of total variance) was significantly explained by genus affiliation ($P < 0.001$). Species affiliation accounted for 36% of the total variance of ERIC incidence ($P < 0.001$). Variance due to strain-to-strain difference was limited to 5% of the total variance ($P < 0.001$). When the strain effect was reanalyzed while regarding the high variability in three strains of *D. dadantii* as exceptional outliers, difference in strain was not a significant factor causing variability in the copy number of ERIC sequences in a genome ($P = 0.20$).

Diversity of ERIC sequences

Diversity among ERIC sequences was analyzed as intra-genome diversity and inter-genome divergence along taxonomic hierarchy. Nucleotide diversity values ranged from 0.02 to 0.34 while the majority of genomes had their ERIC diversity values 0.10-0.24 (Fig. 1). As seen from the diversity values

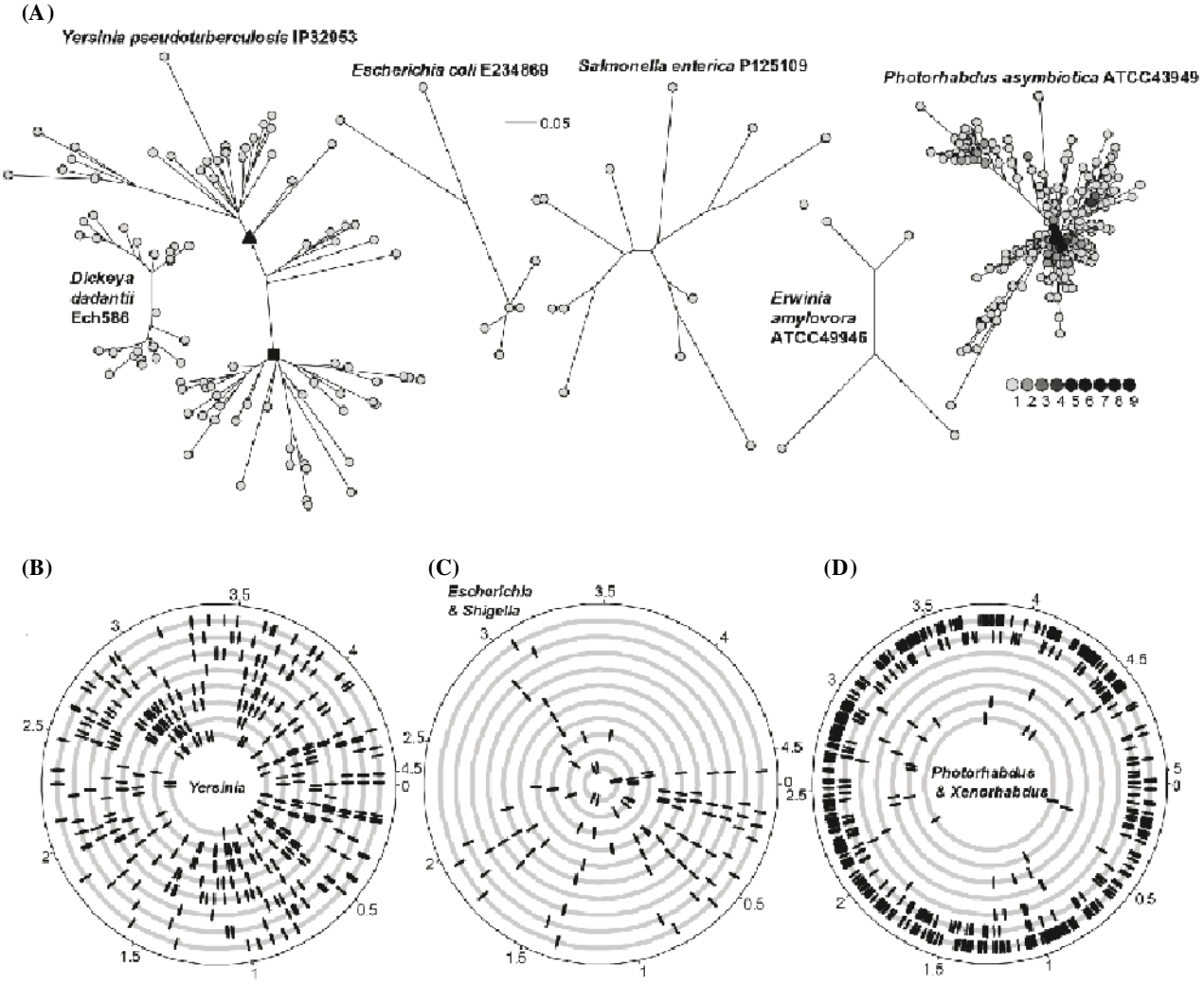

Fig. 2. Diversity and locations of ERIC sequences by genome. (A) Unrooted maximum likelihood trees for selected strains. All trees were drawn with the same scale. Scales: bar = substitutions per base; density in circle (1-9) = the number of each sequence type. Symbols in *Y. pseudotuberculosis* IP 32953 tree (triangle and square): nodes supported by bootstrapping. (B) locations of ERIC loci on circular chromosomes of *Yersinia* strains. Rings: the thin black outermost ring = scale ring, denoting locations (Mbp) from the replication origin; thick gray rings = chromosomes of four *Y. pestis* strains (Z176003, 91001, KIM, and Angola), and three *Y. pseudotuberculosis* strains (IP 31758, PB1/+ and IP 32953), naming from the outermost to the innermost. Spines on the two innermost rings marks location of the two clusters in the ML tree of *Y. pseudotuberculosis* IP 32953 strain (triangle and square in panel A). The cluster marked with the triangle was shown as spines on the innermost ring. (C) Locations in *Escherichia* and *Shigella* strains. Gray rings (numbered 1-10 from the outermost to the innermost): 1, *E. coli* K-12 W3110; 2, *E. coli* K-12 DH10B; 3, *E. coli* BL21(DE3) Wien; 4, *E. coli* BL21(DE3) KRIBB; 5, *E. coli* ATCC 8739; 6, *E. coli* DH1; 7, *E. fergusonii* ATCC 35469^T; 8, *S. boydii* BS512; 9, *S. flexneri* 301; 10, *S. sonnei* Ss046. (D) locations in *Photorhabdus* and *Xenorhabdus* strains. Gray rings (numbered 1-7 from the outermost to the innermost): 1, all 366 ERIC loci in *P. asymbiotica* ATCC 43949; 2, *P. luminescens* TTO1; 3, *X. nematophila* ATCC 19061; 4-7, locations of one of the three sequence types found at >5 loci on the chromosome of *P. asymbiotica* ATCC 43949.

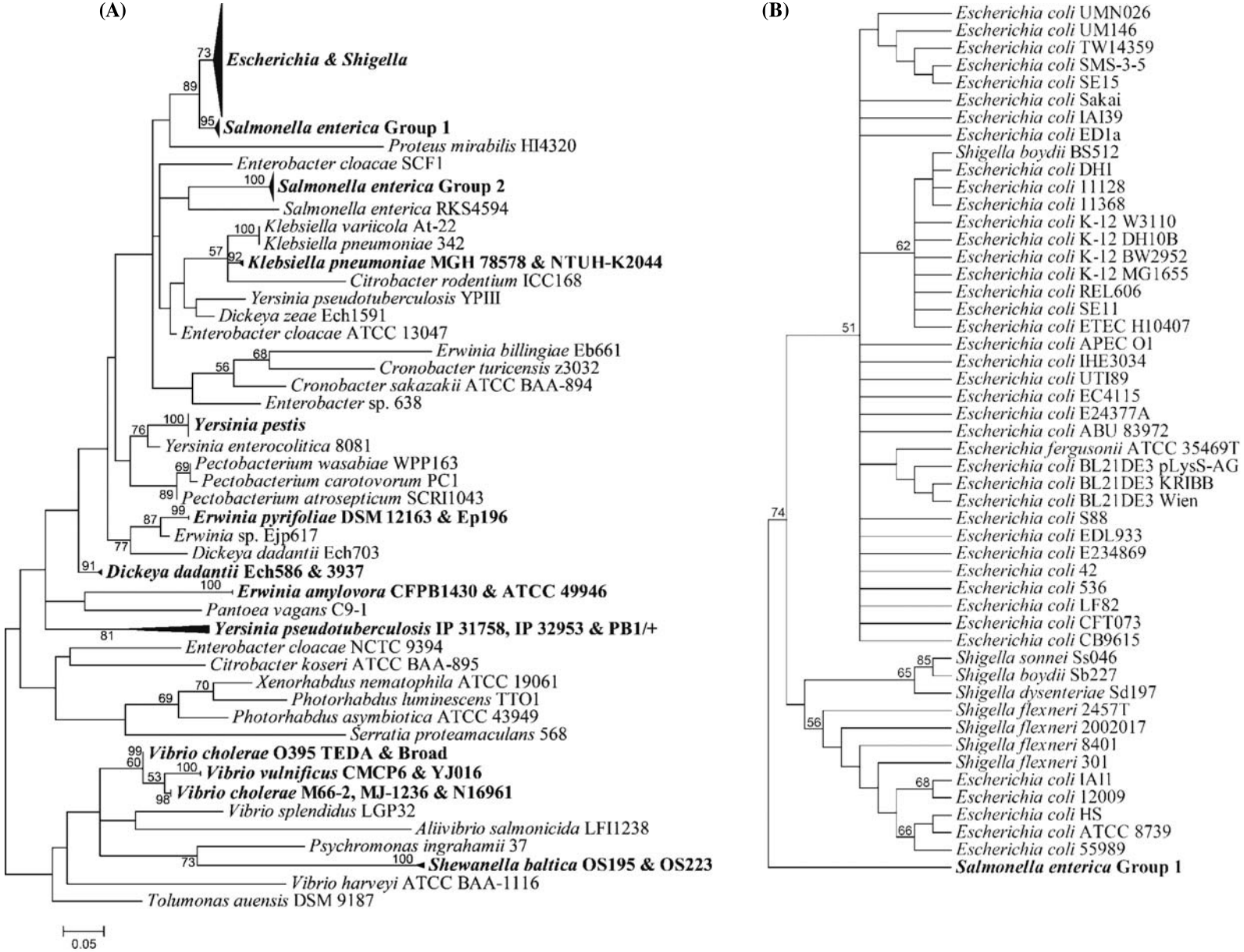

Fig. 3. Phylogenetic relationships among centroids of ERIC sequences found in each genome. (A) Maximum likelihood tree with bootstrap support from 2,000 times of bootstrapping. Genomes of the same species were compressed, except for the cluster of *Escherichia* and *Shigella* genomes. (B) Topology among centroid ERIC sequences of *Escherichia* and *Shigella* genomes, shown as a compressed cluster in panel A.

Table 2. Estimates from analysis of molecular variance (AMOVA)

Structure	Source of variation	df	Sum of squares	Variance	%Variance	*F*-statistic	*P*
STR	Among strain of two families	1	1702	23.3	39.5	F_{CT} 0.40	<0.022
	Among genomes within a strain	6	81	0.0	0.0	F_{SC} 0.00	>0.999
	Within a genome	234	8186	35.7	60.5	F_{ST} 0.40	<0.001
SPP	Among species of three families	13	24923	14.6	24.1	F_{CT} 0.24	<0.001
	Among genomes within a species	94	1410	0.0	0.0	F_{SC} 0.00	>0.999
	Within a genome	1989	88140	45.8	75.9	F_{ST} 0.24	<0.001
GEN	Among genera of five families	11	45918	21.3	36.0	F_{CT} 0.36	<0.001
	Among genomes within a genus	94	4611	0.5	0.8	F_{SC} 0.01	<0.001
	Within a genome	2671	99969	37.4	63.2	F_{ST} 0.37	<0.001
FAM	Among five families	4	16175	16.9	25.4	F_{CT} 0.25	<0.001
	Among genomes within a family	128	38940	11.7	17.6	F_{SC} 0.23	<0.001
	Within a genome	2949	112047	38.0	57.1	F_{ST} 0.43	<0.001

of *Erwinia* spp., *Shigella dysenteriae*, and four strains of *E. coli*, chromosomes with a small number of ERIC loci (<5) tended to outlie from that range. Also notable was that the chromosomes of *Photorhabdus* spp., the strains showing highest incidence of ERIC loci, also outlaid from the range. This pattern implied a certain relationship between nucleotide diversity and the number of ERIC loci exists with some rare exceptions.

Relationship between intra-genome diversity and copy number showed notable trends (Fig. 1). Firstly, diversity values remained within a range when enough a number of ERIC loci were found in a genome. Secondly, the number of ERICs in a genome with exponential increments from a few to hundreds of loci. Thirdly, diversity value varied more among the species with smaller number of ERICs. For example, *E. coli*, *Erwinia*, *Enterobacter*, *Salmonella*, *Shigella*, and *Klebsiella* strains had high variability of ERIC nucleotide diversity. Lastly, the number of loci varied more than diversity when the number of ERICs was large. For *Yersinia* spp., *V. cholerae*, *Pectobacterium* spp., and *Photorhabdus* spp., both diversity and the number of loci tended to cluster according to species or genus identity, but intra-genome diversity was more conserved. This finding was somewhat paradoxical because it implied that ERICs were added to or deleted from genomes of a species without affecting information content. Therefore, it called for careful examination of sequence differences of ERICs in genome of the same species or genus.

When intra-chromosomal phylogeny among ERIC sequences was analyzed by each chromosome, the trend was better understood (see representative ML trees for selected strains in Fig. 2A). Among *E. coli*, *Salmonella*, and *Klebsiella* genomes, most ERIC sequences were distantly related, except for a few clusters made of two sequences. In the chromosomes of *Yersinia* spp., which showed high diversity among large numbers of ERIC loci (Fig. 1), ERIC sequence types were split into two or more clusters (supported by bootstrapping). In *P. asymbiotica* ATCC 43949 and *P. luminescens* TTO1, numerous sets of ERIC loci had either identical or nearly identical sequences, forming clusters of diminished diversity around three more or more core sequences. However, ERIC sequences of those strains were also diverse due to a few highly diverged sequences. ERIC sequences in *Dickeya dadantii* Ech586 showed a small diversity, comparable to those of *Photorhabdus* spp., in spite of small number of ERIC loci. In the *Dickeya* genome, highly diverged sequences were not present, so that all evolutionary distances among ERIC sequences were relatively short.

Inter-genome divergence of ERIC sequences was examined based on phylogeny of the centroid ERIC sequences of each genome and AMOVA of all ERIC sequences along taxonomic hierarchy. Based on bootstrap support value, the ML phylogeny of centroids was characterized as clustering of sequences by genus (Fig. 3A). Deviations from this generalization included separation of strains of an identical genus or species into two separate clusters, e.g., *S. enterica*, *Y. pseudotuberculosis*, and *Enterobacter*. Another deviation was for *Shigella* spp. (Fig. 3B), as it was expected from taxonomic studies stating *Shigella* being a paraphyletic synonym of *E. coli* (Yang *et al.*, 2007). In AMOVA results, intra-genome variation of ERIC sequences accounted for the majority (57%) of the total sequence variation (Structure FAM in Table 2). The difference in family and/or order was a significant factor explaining 25.4% of nucleotide differences in the entire collection of ERIC sequences. The contribution of genus difference was separately estimated based on genera with multiple species or strains (Structure GEN). Likewise, effects of differences in species and strain were estimated by making the top-hierarchy of an AMOVA structure confounded with upper taxonomic hierarchy (Structures STR and SPP). According to the results, difference of strain or species did not cause significant divergence in ERIC sequences. Difference in genus caused a significant divergence, but the amount of divergence was extremely small, i.e., 0.8% of total sequence variation among ERIC sequences from genomes of genera with multiple species or strains. Difference in genera within a given family accounted for about 10% of the total ERIC sequence variation, which was obtained by subtracting the contribution of family (25.4% in Structure FAM) from that of genus (36% in Structure GEN). Bases on these results, it can be concluded that gammaproteobacteria of different genera tends to have different ERIC sequences that are different to each other by either 10% or 36% depending on identity of families which they belong to. On the contrary, bacterial genomes of the same genus are not notably different to each other, but the divergence of ERICs due to genus differences are entangled into the high intra-genome diversity of ERIC sequences.

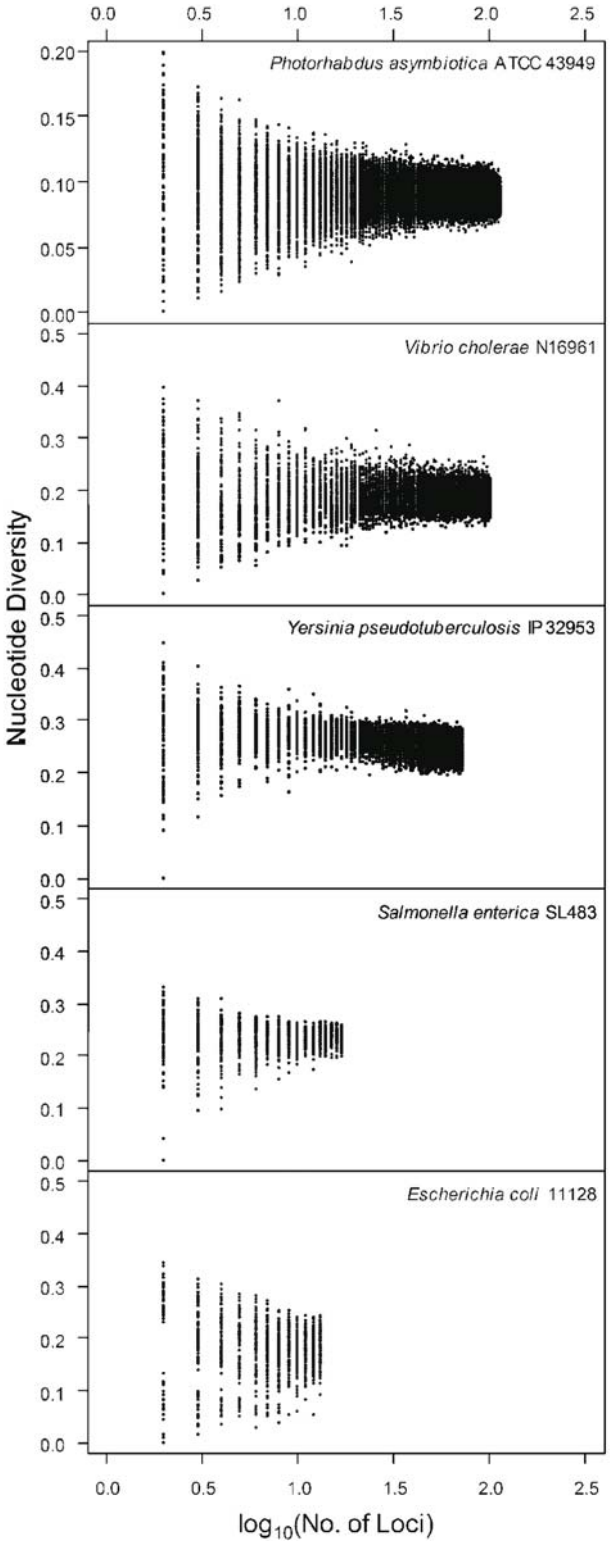

Fig. 4. Simulation of deletion events (100 Jackknifing per a subset size).

Variation in location of ERIC loci

Until today, how ERICs locate and transpose to target site is not clearly understood. For the purpose of elucidating the mechanism(s) responsible for propagation of ERIC sequences up to 366 copies per genome, characteristics in positing of ERIC loci in genomes were analyzed by intra-genome and inter-genome comparisons. To examine whether duplication of an ERIC sequence has tendency of occurring to the vicinity to the source copy, significance was tested for correlation between: the genetic distance value (bp) between a given pair of ERIC sequences in a genome and the physical distance value (bp) between the two ERIC loci. Spearman rank correlation coefficient was used as the estimate for correlation among pairwise distances, and the significance was tested by Mantel's test with 20,000 permutations. For 114 out of 127 genomes with ERIC sequence, no significant correlation was found between genetic distances and physical distances of ERIC loci ($P>0.05$; Mantel's test). Eleven out of the other 13 genomes with significant P value in Mantel's test included genomes of *Y. pseudotuberculosis*, *V. cholerae*, *D. dadantii*, *S. enterica*, and *V. vulnificus* strains showing weak correlations ($\rho \leq 0.3$). When physical map of ERIC loci forming a bootstrap supported cluster was examined for each of those strains, the source of the weak correlations turned out to be a large-scale asymmetry, instead of tight clustering of locations. For example, ERICs in *Y. pseudotuberculosis* IP 32953 formed two distinct phylogenetic clusters (the clusters marked with a triangle and a square in Fig. 2A). Sequences belonging to one of the clusters were located asymmetrically along the circular chromosome of the strain (spines on the inner most ring in Fig. 2B). The genomes of two *Erwinia pyrifoliae* strains had strong correlation ($\rho=0.7$) which was significant by Mantel's test ($P=0.01$). However, the fact that the *E. pyrifoliae* genomes had only 5 ERIC loci, those statistics could not be considered as a strong exception to the general trend of no notable correlation between genetic and physical distances among the other 125 genomes. Because the genome of *P. asymbiotica* ATCC 43949 harbored the most number of ERIC loci, it carried several ERIC sequences in perfect replications. Using subsets of ERIC loci each of which had 6-9 replicates of the identical ERIC sequence, spatial pattern of ERIC loci propagation was examined in another perspective, under the assumption that one of ERICs in a set of ERIC loci with zero genetic distance value was the first progenitor of all ERICs of the set. As shown in the four innermost rings in Fig. 2D, replicates of identical sequences were distributed throughout the genome rather than being clustered around a focal locus. Therefore, it could be concluded that ERIC loci with similar or identical sequences of nucleotides did not occur in proximity to each other. Rather than being clustered, progeny of an ERIC located randomly throughout the genome.

Stability/variability of ERIC loci in bacterial genomes were tested by examining conservation of locations of ERICs in different genomes of bacteria along their taxonomic hierarchy. The spatial correlation estimate varied widely from zero to unit among pairs of chromosomes. However, the trend of stronger correlation between strains with closer taxonomic affiliation was obvious. For example, ERIC loci on genomes of *Yersinia*, *Escherichia* and *Photorhabdus* were located mostly at different positions (Figs. 2B-D). However, different species of the same genus or different strains of the same species showed highly correlated positioning of ERIC loci. Mean values of circular correlations among all possible pairs of chromosomes were

0.7 for the same strain, 0.5 for the same species, 0.4 for the same genus and 0.2 for the same family. This pattern implied that the locations of ERIC are relatively stable among strains belong to the same genus or species. To formalize this hypothesis, a nested mixed-model ANOVA was performed on the circular correlation coefficient. According to the results, intra-strain variation of spatial correlation coefficient accounted for 61% of its total variance. The rest 39% variation could be split into contributions of genus (13%), species (17%) and strain (9%) with significance ($P<0.05$; LR test). Supra-genus structures such as family and order were not significant explanatory factor for spatial correlation of ERIC loci among genomes. Therefore, it was concluded that a small portion (30%) of chromosomal locations of ERIC loci are conserved in a species-specific manner while a large portion (61%) varies even within a strain.

Discussion

Propagation within a genome

Depending on genus and species affiliation of a bacterium, the number of ERIC loci in a genome varied widely from zero to 366 in gammaproteobacteria. As implied from in Fig. 1 and ANOVA results, the numbers were distributed with exponential increments, and their distribution could be normalized by logarithmic transformation. These numerical characteristics imply that ERIC sequences propagate or vanish within a bacterial lineage in an exponential manner. On the other hand, ERICs of *P. asymbiotica* ATCC 43949 demonstrated that ERICs with different genotypes can be distributed unevenly, ranging from 1 to 9 loci per genotype, with a pattern of point mutations radiating from multiple focal genotypes (Fig. 2A). Also notable was that the set of genotypes radiating from the same focal genotypes have varying number of point mutations in comparison to the focal genotypes. These notions implied that propagation of ERICs in a genome was not a sing exponential propagation, but accumulation of several separate propagation events. ERIC is regarded as a Mariner-like DNA TE, which is mobile by cut-and-paste mode (De Gregorio *et al.*, 2005; Delihas, 2008; Treangen *et al.*, 2009). Therefore, duplicative transpositions are limited to the time of chromosome replication (Muñoz-López and García-Pérez, 2010). Therefore, accumulating the large number of loci will require multiple master copies and many events of duplication for considerable amount time. In the genome of *P. luminescens* TTO1, which did not have a high-frequency genotype, showed 145 genotypes with nearly identical relatives (data not shown), indicating that duplication of ERICs had been intervened by radiating diversification. *Dickeya dadantii* strains also showed clustering of nearly identical genotypes (Fig. 2A). Therefore, the exponential distribution in the copy number of ERICs among gammaproteobacterial lineage could be made either by accumulation of rare incidental propagation events or exponential deletions. However, radiations of highly similar ERIC sequences were not clearly visible among the rest of the genomes examined in this study. Therefore, intra-chromosomal propagation of ERIC sequence appears to be ceased in most of gammaproteobacteria. Active propagation of ERIC sequences appears to be rare incidences confined only in a certain lineage of bacteria.

Another aspect of ERIC sequence propagation examined in this study was whether the propagation occurred to the vicinity to the original copy or not. According to observations on the locations of the high frequency genotypes on the *P. asymbiotica* ATCC 43949 genome (Fig. 2D) and lack of correlation between sequence similarity and physical proximity, duplicative transposition of an ERICs occurs to any location in a chromosome.

Sequence divergence

According to AMOVA results, more than half (57%) of the total variation in ERIC sequences among 129 genomes of Gammaproteobacteria was accounted for by sequence divergence within a genome. On the contrary, number of ERIC sequences was relatively stable among genomes of the same strain. This uncoupling between variations of number and diversity of ERIC nucleotide sequences within a bacterium indicates that they diverged in a genome primarily by an endogenous forcing rather than by uptake of foreign ERIC sequences, i.e., horizontal gene transfer (HGT). For the forcing to be endogenous and universal to all bacteria, random mutation being the forcing appears to be the best explanation.

According to AMOVA results of Structure SPP (Table 2), ERIC sequences found in different genomes of the same species were not diverged significantly, implying that any ERIC sequence on any genome of a species are a members of a common population of ERIC sequences. Structure GEN described only a small proportion of total genetic divergence due to difference in genus (AMOVA, 0.8% $P<0.001$). These phenomena were illustrated in Fig. 2A as similar tree spans, i.e., the maximum of pairwise distances. On the contrary, there was marked variations in the number of ERICs by differences in species and genera (ANOVA, 36% of variance if genera were the same, 36%+45% if genera were different). How did the uncoupling occur? To answer this question, one needs to explain the unknown process changing the number of ERIC sequences in a genome without changing nucleotide sequence content. With propagation or HGT, the number of ERICs will increase. However, the diversity will either decrease for propagation or increase for HGT. When ERIC loci are deleted from a given genome, the number of ERIC loci will decrease without a change in sequence divergence. However, nucleotide diversity value will have larger variance as ERIC loci remaining are a subset of original population of ERIC loci present before the deletion events. When change in nucleotide diversity during deletion events was simulated by jackknifing (Fig. 4), predicted nucleotide diversity values showed larger dispersion as the number of ERIC loci gets smaller, e.g., less than 10, and the overall pattern was similar to Fig. 1. In addition, it was also notable that jackknifed sets of ERIC sequences of *S. enterica* and *E. coli* genomes, which had relatively small number of ERIC loci, were likely to have extreme nucleotide diversity values. This phenomenon could be interpreted as results of genetic drifts from a small population. Based on these interpretations, deletion and genetic drifts of ERIC loci were concluded to be the predominant mechanism for the observed distribution of ERICs.

Reductive evolution

A question subsequent to finding of the importance of dele-

tions of ERICs is why loss of ERIC loci was a widespread process during the evolution of gammaproteobacteria. Because the deletion events appears to occur among all species, negative selection against ERIC sequence is likely to be a universal, constitutive, endogenous mechanism. From noting that ERIC sequence is a stable palindrome structure as DNA or RNA in intergenic regions including those in the middle of polycistronic operons, negative selection against transcription interference can be suggested as the selective pressure for deletions (Mourier and Willerslev, 2008). Regarding that species-specific phylogenetic features of RPEs were also made by loss of vertically inherited RPE loci (Amiri *et al.*, 2002), the tendency of reductive evolution for palindromic genetic elements could be a general pattern in achieving stability of bacterial genomes (Achaz *et al.*, 2003; Treangen *et al.*, 2009). Some of cellular and epigenetic mechanisms checking the activity of retroelements in eukaryotes (Maksakova *et al.*, 2008) might be applicable to bacteria. Also, frequent niche switching from free-living to symbiosis or parasitism, or vice versa, can be considered as a driving force for loss of repeated sequences from gammabacterial genome (Frank *et al.*, 2002; Moran, 2003).

Origins and differential rate of loss

In the BLAST search of this study, ERIC sequences were found only among gammaproteobacteria. Therefore, the origin of ERIC sequences likely to be limited to one of the common progenitors in gammaproteobacteria. In the AMOVA analysis, ERIC sequences of different families of gammaproteobacteria were diverged enough, so that 25% of nucleotide variation in bacteria kingdom was due to differences in five families (Structure FAM in Table 2). Further corroboration for this remark was made by preservation of genus-level phylogeny in ML tree analysis of centroid sequences of ERICs (Fig. 3).

However, vertical transmission alone can be the full explanation of the phylogeny of ERICs in gammaproteobacteria. Class 2 transposable elements are known for its uncanny ability to be transmitted horizontally in many eukaryotic species (Lohe *et al.*, 1995; Lampe *et al.*, 2003; Casse *et al.*, 2006). Therefore, sequence divergence of ERICs that does not conforming the known species phylogeny of gammaproteobacteria can be interpreted as HGT. The cases of *S. enterica*, *Y. pseudotuberculosis*, and *Dickeya* species could be the case (Figs. 2A and 3A). *S. enterica* and *Y. pseudotuberculosis* trees illustrated presence of more than one lineage of ERICs in a genome. In the case of *Dickeya* species, the number of ERIC loci varied most with the range of 6-45. Location of ERIC loci also varied most so that spatial correlation was not significant among *D. dadantii* strains and between *D. dadantii* and *D. zeae*. Furthermore, *D. dadantii* Ech586 strain showed relatively small radiating divergence (Fig. 2A), which indicated relatively recent propagation of ERIC sequences originating from a single master copy introduced by HGT.

While the vertical transmission was primarily modulated by deletion of ERIC loci, the rate of deletion did not appear to be equal for all genus or species. Based on the range of the number of ERIC loci in different genera (Fig. 1), *Photorhabdus*, *Yersinia*, *Pectobacterium*, and *Dickeya* showed high density of ERIC loci while *Escherichia*, *Salmonella*, and *Klebsiella* showed low density. Interestingly, the difference between the two group were ecological because the latter was enteric to warm-blood animals while the former was known as insects/nematodes parasites or soil bacteria suspected to have some ecological relationship with insects/nematodes. Outside the *Enterobacteriaceae* family, *V. cholerae* was the only species with high density of ERIC loci. *V. cholerae* was well known for its commensal relationship with crustacean arthropods (Zo *et al.*, 2008, 2009). Therefore, some unidentified ecological relationships of those species interacting with arthropod/nematode may be the reason for weak negative selection against ERICs or occasional propagation/HGT of ERICs. Regarding the case of *P. asymbiotica*, which showed a clear pattern of recent propagation, even presence of a positive selection for a certain arthropod/nematode-ERIC relationship can be also suggested.

Conclusion

Based on phylogenetic and numerical analyses of this study, the evolutionary path and dynamics of ERICs in bacteria were elucidated. Results of phylogenetic analyses indicated that a mobilized ERIC motif was introduced to a common ancestor of gammaproteobacteria. After cumulative proliferations during replications of ERIC-carrying genomes, interspersed ERIC loci were vertically transmitted under universal constitutive negative selection, forming reductive divergences along the various speciation process of gammaproteobacteria. However, occasional HGT introduced new ERIC loci and/or capacity to propagate in a genome, products of which were also subsequently subjected to the reductive evolution.

Along with those findings, the analyses described in this study also found several useful properties of ERIC sequences in a given genome. Intra-species stability of ERIC loci, in their numbers, sequences and locations, and fixation of divergent sequences by different families suggest ERIC sequences can be used for diagnostics or at least taxonomy applications. Furthermore, the fact that they are under reductive evolution suggest that most of the ERIC loci can be exploited as neutral markers of genome-wide evolution.

Acknowledgements

This research was supported by Kyungsung University Research Grants in 2010.

References

Achaz, G., E. Coissac, P. Netter, and E.P.C. Rocha. 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* 164, 1279-1289.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Amiri, H., C.M. Alsmark, and S.G.E. Andersson. 2002. Proliferation and deterioration of rickettsia palindromic elements. *Mol. Biol. Evol.* 19, 1234-1243.

Bachellier, S., J.M. Clement, and M. Hofnung. 1999. Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol.* 150, 627-639.

Casse, N., Q.T. Bui, V. Nicolas, S. Renault, Y. Bigot, and M. Laulier. 2006. Species sympatry and horizontal transfers of Mariner transposons in marine crustacean genomes. *Mol. Phylog. Evol.* 40, 609-

619.

Cordaux, R. and M.A. Batzer. 2009. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691-703.

De Gregorio, E., G. Silvestro, M. Petrillo, M.S. Carlomagno, and P.P. Di Nocera. 2005. Enterobacterial repetitive intergenic consensus sequence repeats in *Yersiniae*: genomic organization and functional properties. *J. Bacteriol.* 187, 7945-7954.

Delihas, N. 2007. Enterobacterial small mobile sequences carry open reading frames and are found intragenically-evolutionary implications for formation of new peptides. *Gene Regul. Syst. Bio.* 1, 191-205.

Delihas, N. 2008. Small mobile sequences in bacteria display diverse structure/function motifs. *Mol. Microbiol.* 67, 475-481.

Excoffier, L. and H.E.L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564-567.

Excoffier, L., P.E. Smouse, and J.M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479-491.

Frank, A.C., H. Amiri, and S.G. Andersson. 2002. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica* 115, 1-12.

Guindon, S., J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307-321.

Hasegawa, M., H. Kishino, and T.-a. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160-174.

Jammalamadaka, S.R. and A. SenGupta. 2001. Topics in Circular Statistics. World Scientific Press, Singapore.

Lampe, D.J., D.J. Witherspoon, F.N. Soto-Adames, and H.M. Robertson. 2003. Recent horizontal transfer of Mellifera Subfamily *Mariner* transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. *Mol. Biol. Evol.* 20, 554-562.

Lewis, S., E. Akgun, and M. Jasin. 1999. Palindromic DNA and genome stability. Further studies. *Ann. N. Y. Acad. Sci.* 870, 45-57.

Lohe, A.R., E.N. Moriyama, D.A. Lidholm, and D.L. Hartl. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Mol. Biol. Evol.* 12, 62-72.

Maksakova, I.A., D.L. Mager, and D. Reiss. 2008. Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. *Cell. Mol. Life Sci.* 65, 3329-3347.

Moran, N.A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr. Opin. Microbiol.* 6, 512-518.

Mourier, T. and E. Willerslev. 2008. Does selection against transcriptional interference shape retroelement-free regions in mammalian genomes? *PLoS ONE* 3, e3760.

Muñoz-López, M. and J. García-Pérez. 2010. DNA transposons: nature and applications in genomics. *Curr. Genomics* 11, 115-128.

Nei, M. and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York, N.Y., USA.

Nunvar, J., T. Huckova, and I. Licha. 2010. Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. *BMC Genomics* 11, 44.

Ogata, H., S. Audic, V. Barbe, F. Artiguenave, P.E. Fournier, D. Raoult, and J.M. Claverie. 2000. Selfish DNA in protein-coding genes of *Rickettsia*. *Science* 290, 347-350.

Oppenheim, A.B., K.E. Rudd, I. Mendelson, and D. Teff. 1993. Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli*. *Mol. Microbiol.* 10, 113-122.

R Development Core Team. 2005. R: A language and environment for statistical computing. 2.0.1 ed. R Foundation for Statistical Computing, Vienna, Austria.

Tobes, R. and E. Pareja. 2006. Bacterial repetitive extragenic palindromic sequences are DNA targets for insertion sequence elements. *BMC Genomics* 7, 62.

Treangen, T.J., A.L. Abraham, M. Touchon, and E.P. Rocha. 2009. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* 33, 539-571.

Versalovic, J. and J.R. Lupski. 1998. Interspersed repetitive sequences in bacterial genomes. *In* F.J. de Bruijn, J.R. Lupski, and G.M. Weinstock (eds.), Bacterial Genomes: Physical Structure and Analysis, Chapman and Hall, New York, N.Y., USA.

Yang, J., H. Nie, L. Chen, X. Zhang, F. Yang, X. Xu, Y. Zhu, J. Yu, and Q. Jin. 2007. Revisiting the molecular evolutionary history of *Shigella* spp. *J. Mol. Evol.* 64, 71-79.

Zo, Y.G., N. Chokesajjawatee, E. Arakawa, H. Watanabe, A. Huq, and R.R. Colwell. 2008. Covariability of *Vibrio cholerae* microdiversity and environmental parameters. *Appl. Environ. Microbiol.* 74, 2915-2920.

Zo, Y.G., N. Chokesajjawatee, C. Grim, E. Arakawa, H. Watanabe, and R.R. Colwell. 2009. Diversity and seasonality of bioluminescent *Vibrio cholerae* populations in Chesapeake Bay. *Appl. Environ. Microbiol.* 75, 135-146.