The ISME Journal (2011) 5, 1162–1177

www.nature.com/ismej

Open

ORIGINAL ARTICLE

Marine integrons containing novel integrase genes, attachment sites, *attI*, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays

Hosam Elsaied[1,2], Hatch W Stokes[3], Keiko Kitamura[1], Yasurou Kurusu[4], Yoichi Kamagata[1] and Akihiko Maruyama[1]

[1]*Microbial and Genetic Resources Research Group, Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology, Higashi Tsukuba, Ibaraki, Japan;* [2]*Department of Genetics, National Institute of Oceanography, Cairo, Egypt;* [3]*Institute for the Biotechnology of Infectious Diseases, University of Technology, Sydney, New South Wales, Australia and* [4]*Department of Molecular Microbiology, College of Agriculture, Ibaraki University, Ami, Ibaraki, Japan*

In order to understand the structure and biological significance of integrons and associated gene cassettes in marine polluted sediments, metagenomic DNAs were extracted from sites at Suez and Tokyo Bays. PCR amplicons containing new integrase genes, *intI*, linked with novel gene cassettes, were recovered and had sizes from 1.8 to 2.5 kb. This approach uncovered, for the first time, the structure and diversity of both marine integron attachment site, *attI*, and the first gene cassette, the most efficiently expressed integron-associated gene cassette. The recovered 13 and 20 *intI* phylotypes, from Suez and Tokyo Bay samples, respectively, showed a highly divergence, suggesting a difference in integron composition between the sampling sites. Some *intI* phylotypes showed similarity with that from *Geobacter metallireducens*, belonging to Deltaproteobacteria, the dominant class in both sampling sites, as determined by 16S rRNA gene analysis. Thirty distinct families of putative *attI* site, as determined by the presence of an *attI*-like simple site, were recovered. A total of 146 and 68 gene cassettes represented Suez and Tokyo Bay unsaturated cassette pools, respectively. Gene cassettes, including a first cassette, from both sampling sites encoded two novel families of glyoxalase/bleomycin antibiotic-resistance protein. Gene cassettes from Suez Bay encoded proteins similar to haloacid dehalogenases, protein disulfide isomerases and death-on-curing and plasmid maintenance system killer proteins. First gene cassettes from Tokyo Bay encoded a xenobiotic-degrading protein, cardiolipin synthetase, esterase and WD40-like β propeller protein. Many of the first gene cassettes encoded proteins with no ascribable function but some of them were duplicated and possessed signal functional sites, suggesting efficient adaptive functions to their bacterial sources. Thus, each sampling site had a specific profile of integrons and cassette types consistent with the hypothesis that the environment shapes the genome.

The ISME Journal (2011) **5,** 1162–1177; doi:10.1038/ismej.2010.208; published online 20 January 2011
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: integrons; gene cassettes; metagenome; polluted marine sediment

Introduction

Integrons are DNA elements first described in multidrug-resistant pathogenic bacteria (Stokes and Hall, 1989). Their defining features (Figure 1) are an integrase gene, *intI*, which encodes a site-specific recombinase, IntI, and an integron-associated recognition site designated *attI*. This recombination system is designed to capture individual genes when such genes are part of a mobilizable genetic element known as gene cassette (Hall *et al.*, 1999). Insertion of a cassette into an integron occurs via an IntI-mediated site-specific recombination reaction between *attI* and the cassette-associated recombination site, *attC* (also designated a 59-base element in earlier literature) (Collis *et al.*, 2002). Genes in cassettes normally do not include a promoter. However, where examined, integrons possess a promoter, P_c, that drives the transcription of cassette genes when cassettes are inserted in an integron (Collis and Hall, 1995).

The integron recombination system is unusual in comparison to other such systems. Most notable among the differences is the fact that the *attC* sites recognized by the IntI protein are relatively diverse in that such sites are variable in both sequence and length when different gene cassettes are compared.

Correspondence: A Maruyama, Microbial and Genetic Resources Research Group, National Institute of Advanced Industrial, Science and Technology, AIST, 1-1-1, Higashi Tsukuba, Ibaraki 305-8566, Japan.
E-mail: maruyama-aki@aist.go.jp
Received 22 June 2010; revised 14 October 2010; accepted 14 December 2010; published online 20 January 2011

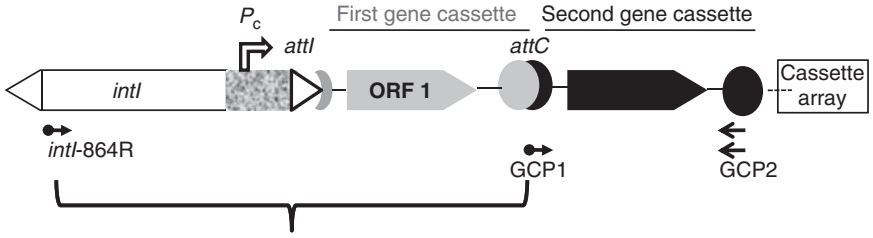


Figure 1 A generalized structure of the marine integrons, indicating *intI* and *attC* specific primer binding sites, marked by left and right arrows, and the sequence region/s missed in previous studies, marked by a brace, and uncovered in this study. *intI*, integrase gene; P_c, gene cassette expression promoter; *attI*, gene cassette attachment site; *attC*, gene cassette base element.

They have a structural conservation, however, and this is important in IntI recognition and reaction catalysis (Stokes *et al.*, 1997; MacDonald *et al.*, 2006). This structural similarity includes the presence of two IntI-like simple sites and an overall structure that comprises an imperfect inverted repeat. The sequence of the *attI* sites within an integron class is highly conserved. However, these sites, like the associated *intI* genes themselves are highly divergent in DNA sequence between classes. One of the most studied *attI* sites is *attI1*, the attachment site associated with class 1 integrons (Partridge *et al.*, 2000). This site comprises a single *intI*-like simple site adjacent to which are two direct repeats, which comprise IntI1-binding domains. *attI* sites of other integron classes possess a simple site but the presence of adjacent direct repeats is more variable (Nield *et al.*, 2001).

Integrons are phylogenetically diverse elements and are dispersed among a wide range of both pathogenic and environmental bacteria (Mazel, 2006; Boucher *et al.*, 2007). At least three integron classes contribute to the spread of antibiotic resistance (Partridge *et al.*, 2009). Outside the antibiotic-resistance domain, the biological significance of environmental integrons was demonstrated in their flexibility to acquire and express adaptive genes for several environmental stresses (Elsaied *et al.*, 2007; Robinson *et al.*, 2007; Koenig *et al.*, 2009; Elsaied and Maruyama, 2010). This unique integron feature creates a system that provides for an enormous pool of adaptive genes to be mobilized, rearranged and disseminated among environmental bacteria. Indeed, the reservoir of adaptive genes capable of being mobilized by integrons has as yet no known upper limit but must, at the very least, number in the thousands. As a result, integrons have a key role in the evolution of bacterial adaptive genome and can greatly influence bacterial diversity and adaptation in ways that other DNA mechanisms cannot (Holmes *et al.*, 2003).

Marine integron/gene cassette metagenomes have been studied based on random PCR amplifications of the *intI* gene, and of gene cassettes in separate techniques, using primers specific for each target (Elsaied *et al.*, 2007; Koenig *et al.*, 2008; Wright *et al.*, 2008; Rodríguez-Minguela *et al.*, 2009). These molecular approaches lacked the ability to show the linkage between *intI* and gene cassettes in a single amplicon. In some cases, this made it difficult to identify integron-associated gene cassettes unambiguously. Moreover, these studies did not recognize the structure and diversity of marine integron *attI* site and integron-associated first gene cassette. The gene cassette recombination site *attC* has been studied in marine integrons from the views of both structure and recombination activity (Elsaied *et al.*, 2007). Consequently, the sequence structure and diversity of *attI* are needed to complete our understanding about the mechanism of integration of marine integron-associated gene cassettes. The first gene cassette, which is located adjacent to the *attI* site of an integron (Figure 1), is the last one integrated and, consequently, involves in the last adaptation response to changing conditions. As it is the closest gene to the promoter, P_c, its expression level is the highest among other integron-associated gene cassettes (Collis and Hall, 1995). Thus, this gene cassette is a good target to find new strongly expressed adaptive genes in metagenomes (Huang *et al.*, 2009).

This study had two aims. The first one was filling of the gap in the structure of the marine integrons by production of PCR amplicons, which contained integron sequences that previously not uncovered, such as P_c, *attI* and the gene cassettes most closely linked to *attI*. The second was demonstrating the functional diversity of marine integrons in capturing gene cassettes, especially the first gene cassette, capable of adaptation and sustaining of bacteria in marine sediments polluted with varieties of industrial wastes. These goals were achieved in metagenomic DNAs extracted from two sampling distant sites, at Suez Bay and Tokyo Bay; with different pollutant characteristics that may enable in understanding the hypothesis that integrons have the potential to capture genes important in bacterial niche adaptation. We combined both statistical and phylogenetic analyses to estimate the bacterial communities, based on the 16S rRNA gene and composition of *intI* in the sampling sites.

Materials and methods

Sampling

Grab samples of surface sediments were collected at water depths of about 20 m from two Bay sites,

where the pollution stresses were expected to stimulate spreading and diversity of integron/gene cassette metagenome (Koenig *et al.*, 2009). The first site, El-Zeitia, is located at the north terminus of Suez Bay and constitutes the south gate of Suez Canal, Egypt, 29° 57.160′ N, 32° 31.725′ E. El-Zeitia is the most contaminated site of Suez Bay. It is exposed daily to high loads of petroleum wastes coming from the petroleum refining industry around the site, crude oil spills from the fixed oil pipes under the water and ships traveling through the Suez canal as well as other various anthropogenic wastes from the surrounding urban region (El-Agroudy *et al.*, 2006; Nemr *et al.*, 2006; this study) (Supplementary Table 1). The second site is located at the center of Tokyo Bay, 35° 26.71′ N, 139° 50.18′ E, where the sediment is contaminated with a mixture of industrial domestic wastes such as chlorinated polycyclic aromatic hydrocarbons such as dioxin and perfluorinated compounds (Naito and Murata, 2007; Uchimiya *et al.*, 2007; Horii *et al.*, 2009; Zushi *et al.*, 2010) (Supplementary Table 1). The sediment samples were treated with 100 mM EDTA buffer and kept at −80° C for further analyses.

Metagenomic DNA extraction and molecular analyses
Bulk microbial DNAs were extracted from 50 g of sediment, 10 g per each extraction, at each sampling site using PowerMax Soil DNA Isolation Kit (Catalog no. 12988-10, Mo Bio Laboratories, Carlsbad, CA, USA) according to the manufacture's protocol with modifications. A combination of mechanical force using beads, heat and chemical detergents was applied to lyse microbial sediment cells. The released crude lysates were bound to silica spin filters to purify metagenomic DNAs. The sizes of the extracted DNAs were checked by electrophoresis on a 0.9% agarose gel against a Lambda-*Hin*dIII digest marker (New England BioLabs, Hitchin, Hertfordshire, UK) with ethidium bromide staining.

To generate PCR amplicons, containing *intI*, the gene cassette promoter, P_c, the integron attachment site, *attI*, and integrated gene cassettes, the primer *intI*-864R, specific for binding to the sequence encodes the conserved amino acids RHS(T)FATHLL, IntI box II, was used coupling with the primer GCP2, designed from a conserved region of gene cassette base element, *attC*, (Table 1; Figure 1) (Elsaied *et al.*, 2007). PCR reaction mixture, 50 µl, contained 10 × LA buffer II (Mg^{2+} plus), 0.2 µM primer, 400 µM dNTP each, 2.25 U Takara LA Taq Polymerase (Takara, Japan) and 5–30 ng DNA template. Shuttle, combined annealing-extension temperatures, long PCR amplification reaction was performed with the thermal cycler, iCycler (Bio-Rad, Richmond, CA, USA). PCR was started with an initial denaturation at 96 °C for 2 min, followed by 30 cycles with 30 s at 95 °C, annealing extension at 60–62 °C for 10 min, with increasing of 1 °C every 10 cycles, followed by a final elongation at 72 °C for 10 min. PCR products were checked by electrophoresis on a 1.0% agarose gel.

Another PCR approach, using the primers GCP1 and GCP2, was used to recover gene cassettes generally (that is not uniquely the first cassette) from the studied metagenomic DNAs according to the methodology described by Elsaied *et al.* (2007).

In order to identify the bacterial communities at the sampling sites, the primers 27F and 1492R (Table 1) (Lane *et al.*, 1985) were used to amplify the bacterial 16S rRNA gene from the metagenomic DNAs from which the integron amplicons were obtained.

All the PCR products had 3′-A overhangs to facilitate TA-cloning into TOP10 *Escherichia coli* using a TOPO XL PCR-cloning kit according to the manufacturer's instructions (Catalog no. K4750-20, Invitrogen Life Technologies, Carlsbad, CA, USA). Only cells containing XL-TOPO vector with the insert were competent to grow with kanamycin. Colonies of these cells were screened directly by sequencing using vector primers T7 and an ABI 3730 × l 96-capillary DNA analyzer (Applied Biosystems, Foster City, CA, USA).

Sequence analyses
Sequencing results were introduced to FASTA, http://fasta.ddbj.nig.ac.jp/top-e.html, to determine their similarity to known sequences deposited in DNA database. Sequences for *intI* and gene cassette open reading frames (ORFs) were submitted to Transeq, http://www.ebi.ac.uk/emboss/transeq/, to obtain the inferred amino-acid sequences. The correct *intI* ORFs were identified from the presence of diagnostic motifs. Gene cassette promoter

Table 1 Primers used in this study

Primer	*Sequence*	*Primer*	*Expected product size*
intI-864R[a]	5′-YAGCAGATGNGTGGCRAAVSWRTGSCG-3′	GCP2	Varied
GCP1[b]	5′-GCSGCTKANCTCVRRCGTTRRRY-3′	GCP2	Varied
GCP2[b]	5′-TCSGCTKGARCGAMTTGTTRRRY-3′		
27F[c]	5′-AGAGTTTGATCCTGGCTCAG-3′	1492R	~1500 bp
1492R[c]	5′-GGTTACCTTGTTACGACTT-3′		

[a]Primer targets integrase gene *intI*.
[b]Primer targets gene cassette base element *attC*.
[c]Primer targets 16S rRNA gene.

sequence, P_c, was identified by the promoter predictor in the software GENETYX, ver.9, http://www.sdc.co.jp/genetyx/product/genetyx9/news.html. *attI* sequence was identified by observing the simple gene cassette integration site, consisting of a pair of inversely oriented IntI1-binding domains, together with two further directly oriented IntI1-binding sites designated strong and weak (Partridge *et al.*, 2000). Gene cassettes were identified as described previously (Stokes *et al.*, 2001).

Statistical analyses
Grouping of sequences into phylotypes was done, based on genetic distances, using the Mothur software package V.1.7.2 (Schloss *et al.*, 2009), where *intI* and 16S rRNA gene sequences, having 100% deduced amino acid identities and >97.0% nucleotide identities over the regions compared, respectively, were grouped into a single phylotype (Hartmann and Widmer, 2006; Elsaied *et al.*, 2007). The gene cassette richness was determined based on grouping of cassettes that had >70.0% nucleotide identities into a single cassette type (Koenig *et al.*, 2008).

The diversities of the obtained sequences were analyzed by several methods in the Mothur software package. The LIBSHUFF analyses (Singleton *et al.*, 2001) were used to estimate homologous and heterologous coverage of both *intI* and 16S rRNA gene clone libraries as a function of evolutionary distance for pairwise reciprocal comparisons (library A compared with library B and *vice versa*). Differences in coverage were considered significant at *P*-values of <0.05. Rarefaction analyses (Simberloff, 1978) were used to plot the expected number of phylotypes as a function of the number of sequences sampled. Chao 1 richness estimator and Shannon–Weiner index were applied to calculate the potential number and diversity of gene cassette types, respectively (Shannon and Weaver, 1963; Chao, 1984).

Phylogenetic analyses
The phylogenetic analyses, based on deduced amino-acid and nucleotide sequences of the *intI* and 16S rRNA gene phylotypes, respectively, and corresponding sequences from the databases, were performed by applying the neighbor-joining algorithm and drawing the trees using the MEGA 3.1 software (http://www.megasoftware.net/). The branching patterns of the constructed phylogenetic trees were confirmed by reconstruction of the phylogenies using two other methods of analysis, namely maximum parsimony and maximum-likelihood, contained within the Phylip package, http://evolution.genetics.washington.edu/phylip.html.

Protein prediction analyses
First gene cassette ORF deduced proteins that had no significant homology with those in databases were introduced to online protein prediction programs InterProScan sequence search, Psortb and ProtParam, located at ExPASy proteomics server http://www.expasy.org/tools/#secondary, in order to predict protein functional sites, cellular localization, theoretical isoelectric focusing point, pI and stability, respectively.

Nomenclature and accession numbers of the recorded sequences
All recovered sequences were from metagenomic DNAs and consequently, the source organism could not be identified. Hence, we have adopted a nomenclature whereby each clone was the descriptor of the name of the sampling site followed by the number of the clone, for example, Suez1 and so on. The gene cassette had the same site name description followed by two letters GC (Gene Cassette(s)); and two numerical codes representing the number of the clone followed by the number of the gene cassette; for example, SuezGC1.1 means the first gene cassette in Suez clone 1, while SuezGC1.2 means the second gene cassette in the same clone.

The recorded sequences were deposited in the DNA database under accession numbers from AB546978 to AB546998 and from AB547085 to AB547105 for sequences, which contained *intI*, P_c, *attI* and gene cassettes, recovered from Suez and Tokyo Bay samples, respectively. The accession numbers from AB546999 to AB547084 and from AB547106 to AB547123 represented gene cassette sequences generated by the primers GCP1 and GCP2 from Suez and Tokyo Bay samples, respectively. The 16S rRNA gene sequences were deposited under accession numbers from AB530169 to AB530200 and from AB530201 to AB530247 for Suez and Tokyo Bay sequences, respectively.

Results and discussion

Production of diverse PCR amplicons contained all integron structural features
From two different marine sediment metagenomes, we could amplify DNA fragments that contained *intI*, the gene cassette promoter P_c, *attI* and at least the first integrated gene cassette up to the point of the binding of the primer GCP2. Amplicons had a length of up to about 2.5 kb. Sequence analysis of 21 randomly selected clones from each clone library, containing identifiable inserts, implied that most of the clones in the current libraries contained integron-related sequences based on the presence of relevant features. The high efficiency of recovery suggests that the primers used here were more selective than those originally used by Nield *et al.* (2001) where a high number of false positives was observed. The integron-containing amplicons had sizes ranging from 1805 to 2180 bp and from 1771 to 2500 bp, a feature of diverse amplicons obtained from Suez and Tokyo Bay metagenomes,

respectively. Amplicon size differences were predominantly due to the presence of the different length and number of recovered integrated gene cassettes.

The sampling sites were dominated by Delta/Epsilonproteobacteria and unique compositions of integrase

16S rRNA gene sequences were used to identify the bacterial communities and to provide an indication of the putative sources of integrons in the sampling sites. Sequence analyses of 50 clones from each of the 16S rRNA gene clone libraries obtained 32 and 47 phylotypes, representing the bacterial diversity in Suez and Tokyo Bay samples, respectively.

Both samples were dominated by phylotypes belonging to the Delta/Epsilon subdivision of Proteobacteria. Also, both samples contained phylotypes related to Gammaproteobacteria, Planctomycetes and unclassified bacteria. On the other hand, some phylotypes formed unique phylogenetic lineages characterizing each sampling site. The Tokyo Bay sample showed an expansion of bacterial diversity by occurrence of phylotypes belonging to bacterial phyla that were not recorded in the Suez Bay sample such as Bacteroidetes and Chlorobi. In contrast, the phyla Chloroflexi and Firmicutes were represented only in the Suez Bay sample (Figure 2a).

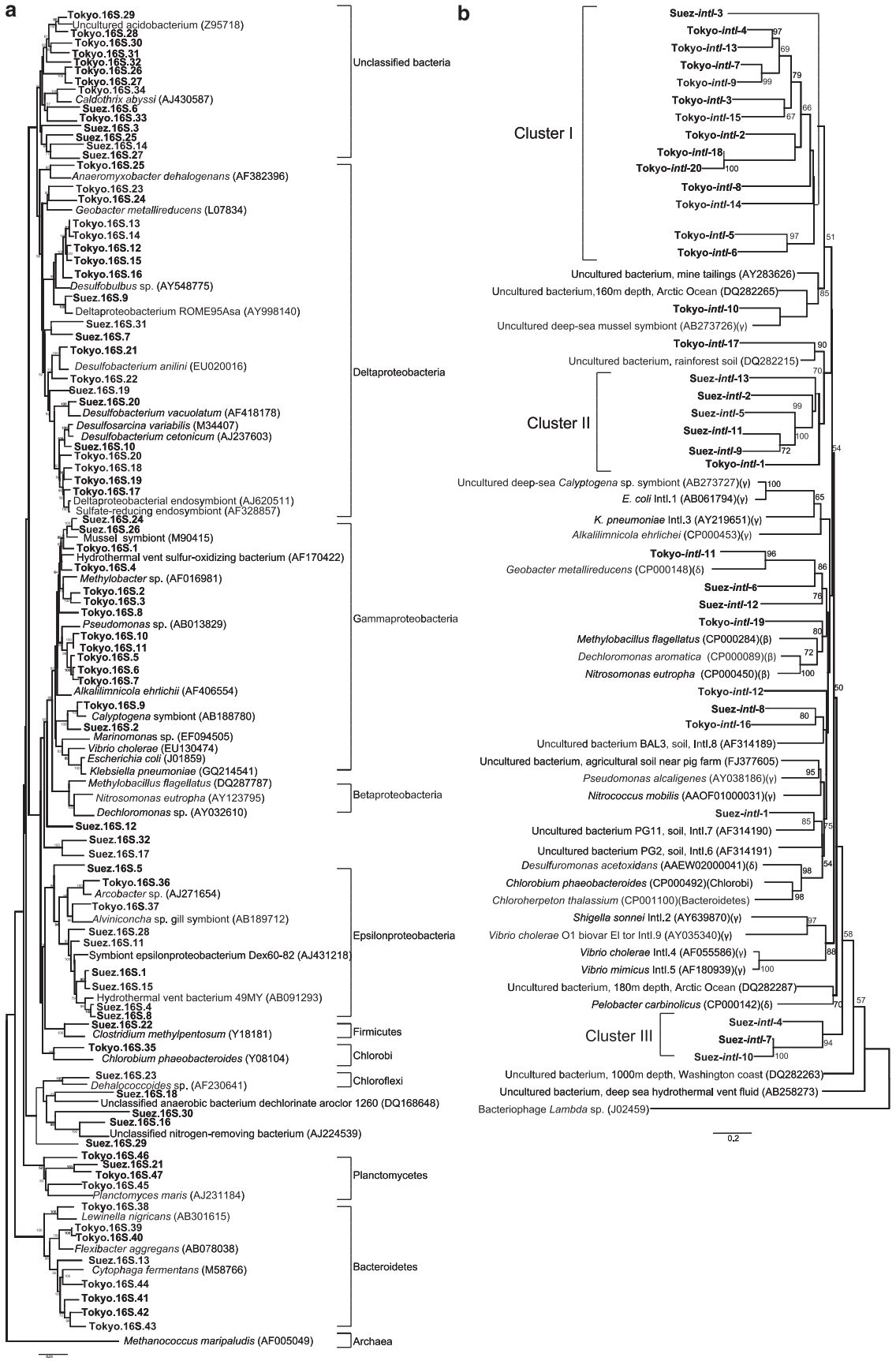
Analyses of 21 integron-containing sequences, from each sampling site, obtained 13 and 20 *intI* phylotypes from Suez and Tokyo Bay samples, respectively. The rarefaction curves for both *intI* and 16S rRNA gene sequences of Tokyo Bay sample showed parallel unsaturated diversities (Figures 3a and b), implying that the analyzed sequences represented a fraction of the total phylotype composition of the site. Although the 16S rRNA gene phylotypes could mostly cover the bacterial composition in the Suez sample, as shown by saturation of the rarefaction curve after only 30 clones, the *intI* rarefaction curve did not show saturation and further sampling of Suez *intI* is likely to have revealed additional diversity. Unsaturated sampling of *intI* has been reported in almost all previously studied marine metagenomes (Rodríguez-Minguela *et al*., 2009). The environments from which these samples were derived were polluted. This was most evident in the Suez Bay sample and this may lead to a limitation in microbial diversity by dominance of only few phylotypes, but with diverse integrons and gene cassettes, which are able to tolerate contamination (Berthe-Corti and Nachtkamp, 2010).

The results from LIBSHUFF analyses showed that the Suez 16S rRNA gene clone library did not differ from that of Tokyo ($P = 0.06$) (Table 2). The only explanation of this pattern was that the diversity of the Tokyo library sequences encompassed and described some sequences in the Suez library. On the other hand, Suez-*intI* phylotypes differed significantly from that of Tokyo, indicating the high divergence between the two *intI* populations. The fact that the site-specific environmental characteristics shape the structure of integrons may establish endemic *intI* pools (Labbate *et al*., 2009). This *intI* sequence divergence was clearly demonstrated in the distribution of most of current *intI* phylotypes within three unique phylogenetic clusters (Figure 2b). Tokyo-*intI* phylotypes forming the cluster I showed phylogenetic distinction from the phylotype Suez-*intI*-3. A similar phylogenetic differentiation occurred between Suez-*intI* phylotypes and Tokyo-*intI*-1 within the cluster II. The cluster III comprised only three Suez-*intI* phylotypes constituting unique phylogenetic lineages. On the other hand, several *intI* phylotypes showed phylogenetic relationships with those recovered from different environmental DNA sources such as deep-sea and soil (Figure 2b), expanding the biogeographically distribution of integrons in both marine and terrestrial environments (Elsaied *et al*., 2007; Rodríguez-Minguela *et al*., 2009). It was not possible to unambiguously identify the organisms that contained these integrons. However, the phylotype Tokyo-*intI*-11 showed phylogenetic relationship, by forming a monophyletic clade, with that of *Geobacter metallireducens*, a species represented by the current phylotypes Tokyo.16S.23 and Tokyo.16S.24, suggesting bacterial sources for integrons in Tokyo Bay sample (Figures 2a and b).

The one previous study that recovered near full-length environmental *intI* sequences was very limited in scope (Nield *et al*., 2001) and other studies have recovered only small fragments internal to the gene (~400 bp) and without associated gene cassette contextual information (Elsaied *et al*., 2007; Rodríguez-Minguela *et al*., 2009). The current study obtained *intI* genes with sizes ranged from 738 to 1269 bp, and from 696 to 1269 bp; from Suez and Tokyo Bay samples, respectively, revealing near complete *intI* sequence structure, including the N-terminus, and contained all IntI motifs including Patch I, box I, Patch II, Patch III and most of box II (Nunes-Duby *et al*., 1998).

Multiple alignments of deduced amino acids of current *intI* phylotypes showed a conservation in

Figure 2 Consensus trees were constructed based on 16S rRNA gene nucleotide sequences (**a**) and *intI* deduced amino-acid sequences (**b**). The trees showed the phylogenetic relationship between the recovered sequence phylotypes and their homologues from database. The trees were constructed by neighbor-joining analyses and confirmed by maximum-parsimony and maximum-likelihood methods. Bootstrap values of >50% were indicated at the branch roots. Recovered sequences were highlighted in bold. Abbreviations (β), (γ) and (δ) represented the classes Betaproteobacteria, Gammaproteobacteria and Deltaproteobacteria, respectively. The bars represented 0.05 and 0.2 changes per a nucleotide and an amino acid, respectively.


a
Tokyo.16S.29
Uncultured acidobacterium (Z95718)
Tokyo.16S.28
Tokyo.16S.30
Tokyo.16S.31
Tokyo.16S.32
Tokyo.16S.26
Tokyo.16S.27
Tokyo.16S.34
Caldothrix abyssi (AJ430587)
Suez.16S.6
Tokyo.16S.33
Suez.16S.3
Suez.16S.25
Suez.16S.14
Suez.16S.27
Unclassified bacteria
Tokyo.16S.25
Anaeromyxobacter dehalogenans (AF382396)
Tokyo.16S.23
Tokyo.16S.24
Geobacter metallireducens (L07834)
Tokyo.16S.13
Tokyo.16S.14
Tokyo.16S.12
Tokyo.16S.15
Tokyo.16S.16
Desulfobulbus sp. (AY548775)
Suez.16S.9
Deltaproteobacterium ROME95Asa (AY998140)
Suez.16S.31
Suez.16S.7
Tokyo.16S.21
Desulfobacterium anilini (EU020016)
Tokyo.16S.22
Suez.16S.19
Suez.16S.20
Desulfobacterium vacuolatum (AF418178)
Desulfosarcina variabilis (M34407)
Desulfobacterium cetonicum (AJ237603)
Suez.16S.10
Tokyo.16S.20
Tokyo.16S.18
Tokyo.16S.19
Tokyo.16S.17
Deltaproteobacterial endosymbiont (AJ620511)
Sulfate-reducing endosymbiont (AF328857)
Deltaproteobacteria
Suez.16S.24
Suez.16S.26
Mussel symbiont (M90415)
Tokyo.16S.1
Hydrothermal vent sulfur-oxidizing bacterium (AF170422)
Tokyo.16S.4
Methylobacter sp. (AF016981)
Tokyo.16S.2
Tokyo.16S.3
Tokyo.16S.8
Pseudomonas sp. (AB013829)
Tokyo.16S.10
Tokyo.16S.11
Tokyo.16S.5
Tokyo.16S.6
Tokyo.16S.7
Alkalilimnicola ehrlichii (AF406554)
Tokyo.16S.9
Calyptogena symbiont (AB188780)
Suez.16S.2
Marinomonas sp. (EF094505)
Vibrio cholerae (EU130474)
Escherichia coli (J01859)
Klebsiella pneumoniae (GQ214541)
Gammaproteobacteria
Methylobacillus flagellatus (DQ287787)
Nitrosomonas eutropha (AY123795)
Dechloromonas sp. (AY032610)
Betaproteobacteria
Suez.16S.12
Suez.16S.32
Suez.16S.17
Suez.16S.5
Tokyo.16S.36
Arcobacter sp. (AJ271654)
Tokyo.16S.37
Alviniconcha sp. gill symbiont (AB189712)
Suez.16S.28
Suez.16S.11
Symbiont epsilonproteobacterium Dex60-82 (AJ431218)
Suez.16S.1
Suez.16S.15
Hydrothermal vent bacterium 49MY (AB091293)
Suez.16S.4
Suez.16S.8
Epsilonproteobacteria
Suez.16S.22
Clostridium methylpentosum (Y18181)
Firmicutes
Tokyo.16S.35
Chlorobium phaeobacteroides (Y08104)
Chlorobi
Suez.16S.23
Dehalococcoides sp. (AF230641)
Chloroflexi
Suez.16S.18
Unclassified anaerobic bacterium dechlorinate aroclor 1260 (DQ168648)
Suez.16S.30
Suez.16S.16
Unclassified nitrogen-removing bacterium (AJ224539)
Suez.16S.29
Tokyo.16S.46
Suez.16S.21
Tokyo.16S.47
Tokyo.16S.45
Planctomyces maris (AJ231184)
Planctomycetes
Tokyo.16S.38
Lewinella nigricans (AB301615)
Tokyo.16S.39
Tokyo.16S.40
Flexibacter aggregans (AB078038)
Suez.16S.13
Cytophaga fermentans (M58766)
Tokyo.16S.44
Tokyo.16S.41
Tokyo.16S.42
Tokyo.16S.43
Bacteroidetes
Methanococcus maripaludis (AF005049)
Archaea
b
Cluster I
Suez-intI-3
Tokyo-intI-4
Tokyo-intI-13
Tokyo-intI-7
Tokyo-intI-9
Tokyo-intI-3
Tokyo-intI-15
Tokyo-intI-2
Tokyo-intI-18
Tokyo-intI-20
Tokyo-intI-8
Tokyo-intI-14
Tokyo-intI-5
Tokyo-intI-6
Uncultured bacterium, mine tailings (AY283626)
Uncultured bacterium,160m depth, Arctic Ocean (DQ282265)
Tokyo-intI-10
Uncultured deep-sea mussel symbiont (AB273726)(γ)
Tokyo-intI-17
Uncultured bacterium, rainforest soil (DQ282215)
Cluster II
Suez-intI-13
Suez-intI-2
Suez-intI-5
Suez-intI-11
Suez-intI-9
Tokyo-intI-1
Uncultured deep-sea Calyptogena sp. symbiont (AB273727)(γ)
E. coli IntI.1 (AB061794)(γ)
K. pneumoniae IntI.3 (AY219651)(γ)
Alkalilimnicola ehrlichei (CP000453)(γ)
Tokyo-intI-11
Geobacter metallireducens (CP000148)(δ)
Suez-intI-6
Suez-intI-12
Tokyo-intI-19
Methylobacillus flagellatus (CP000284)(β)
Dechloromonas aromatica (CP000089)(β)
Nitrosomonas eutropha (CP000450)(β)
Tokyo-intI-12
Suez-intI-8
Tokyo-intI-16
Uncultured bacterium BAL3, soil, IntI.8 (AF314189)
Uncultured bacterium, agricultural soil near pig farm (FJ377605)
Pseudomonas alcaligenes (AY038186)(γ)
Nitrococcus mobilis (AAOF01000031)(γ)
Suez-intI-1
Uncultured bacterium PG11, soil, IntI.7 (AF314190)
Uncultured bacterium PG2, soil, IntI.6 (AF314191)
Desulfuromonas acetoxidans (AAEW02000041)(δ)
Chlorobium phaeobacteroides (CP000492)(Chlorobi)
Chloroherpeton thalassium (CP001100)(Bacteroidetes)
Shigella sonnei IntI.2 (AY639870)(γ)
Vibrio cholerae O1 biovar El tor IntI.9 (AY035340)(γ)
Vibrio cholerae IntI.4 (AF055586)(γ)
Vibrio mimicus IntI.5 (AF180939)(γ)
Uncultured bacterium, 180m depth, Arctic Ocean (DQ282287)
Pelobacter carbinolicus (CP000142)(δ)
Cluster III
Suez-intI-4
Suez-intI-7
Suez-intI-10
Uncultured bacterium, 1000m depth, Washington coast (DQ282263)
Uncultured bacterium, deep sea hydrothermal vent fluid (AB258273)
Bacteriophage Lambda sp. (J02459)
0.2

structural motifs with those from clinical and environmental isolates, but included some amino-acid mutations (Messier and Roy, 2001; Nemergut *et al*., 2004) (Supplementary Figure 1). Several Tokyo-*intI* phylotypes showed substitutions in the conserved glutamic acid, E121, patch I motif. The phylotype Tokyo-*intI*-20 contained four additional amino acids in the patch II motif beside several mutations including substitutions of argenine (R) instead of the conserved lysine (K171), patch II motif; serine (S) and proline (P) instead of the conserved tryptophan (W229) and phenylalanine (F233), respectively, patch III motif; and alanine (A) instead of conserved histidine, H277, box II motif (Supplementary Figure 1). On the other hand, the phylotype Tokyo-*intI*-13 lacked five amino acids including the catalytic amino acid H277, box II motif. It was found that the substitutions in W229 and H277 resulted in IntI1 being completely unable to bind to the *attI1* site, while substitutions in E121 and K171 did not disturb DNA-binding activity of IntI1 (Messier and Roy, 2001). The phylotype Tokyo-*intI*-10, the shortest current IntI sequence, 232 deduced amino acids, suffered from deletions of the motifs box I, patch II, patch III and the first four amino acids, including the catalytic H277, of box II. It was not clear whether this phylotype, Tokyo-*intI*-10, represented an *intI* pseudogene that had lost the protein-coding ability and consequently, recombinase activity. Recent multigenome analyses of bacteria revealed that the proportion of integrase pseudogenes is significantly high, suggesting a different evolutionary dynamic for this class of genes (Liu *et al*., 2004; Nemergut *et al*., 2008). However, this *intI* sequence was located adjacent to *attI* and two integrated gene cassettes with an identifiable *attC* (acc. no. AB547094), representing an integron-associated gene cassette structure.

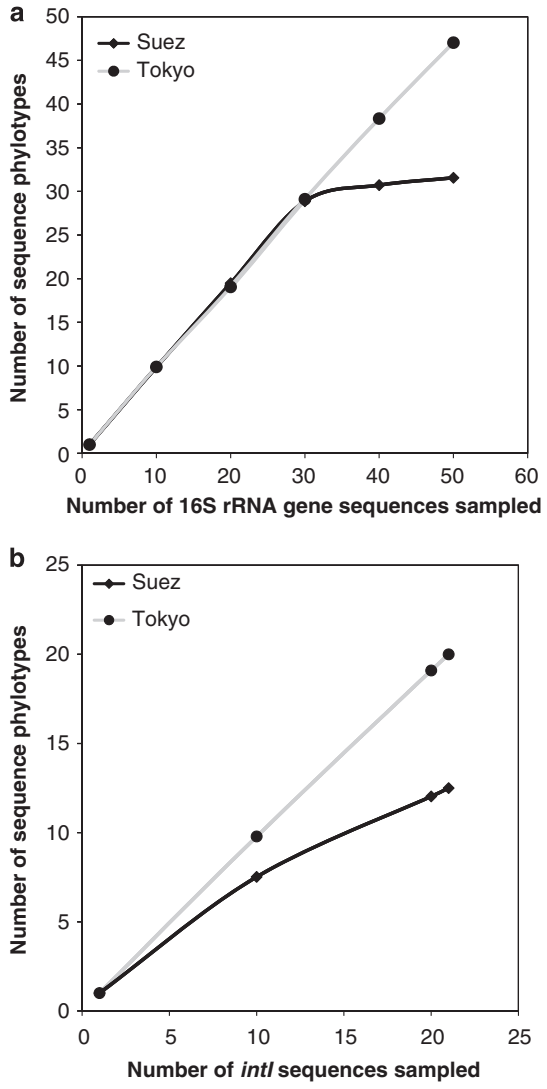


Figure 3 Rarefaction curves for the expected number of 16S rRNA gene phylotypes (**a**) and *intI* phylotypes (**b**).

Recovered putative attI sites were diverse and possessed a characteristic simple site but mostly lacked IntI-like direct repeat binding sites

This study recorded 30 diverse families of *attI*-like sequences from the studied marine integrons (Figure 4). Multiple alignments with *attI1* and those recorded previously in environmental samples showed the existence of the putative *attI* integration simple site, the essential region for *attI* recombination activity, in all recovered sequences (Figure 4). Moreover, the locations of the putative *attI* sites relative to the linked *intI* gene were consistent with known integrons (Partridge *et al*., 2000). The simple site contained known IntI1-binding domains in the clones Suez20, Tokyo5, 9, 10 and 14 (Figure 4).

While possessing an identifiable simple site, most of the recovered *attI* sites in this study lacked

Table 2 LIBSHUFF coverage for integrase and 16S rRNA gene clone libraries

Gene	*Clone library*	*Analyzed clones*	*Phylotypes*	Cov_{hom} *(%)*	Cov_{het} *(%)*	P
Integrase *intI*	Suez	21	13	84.0	60.0	0.001
	Tokyo	21	20	70.0	47.0	0.007
16S rRNA gene	Suez	50	32	56.0	55.0	0.060
	Tokyo	50	47	65.0	52.0	0.001

Homologous (Cov_{hom}) and heterologous (Cov_{het}) coverage percentages of libraries were given. Probability values (*P*) for the significance of differences between homologous and heterologous coverage in reciprocal comparisons as a function of evolutionary distance were also given.

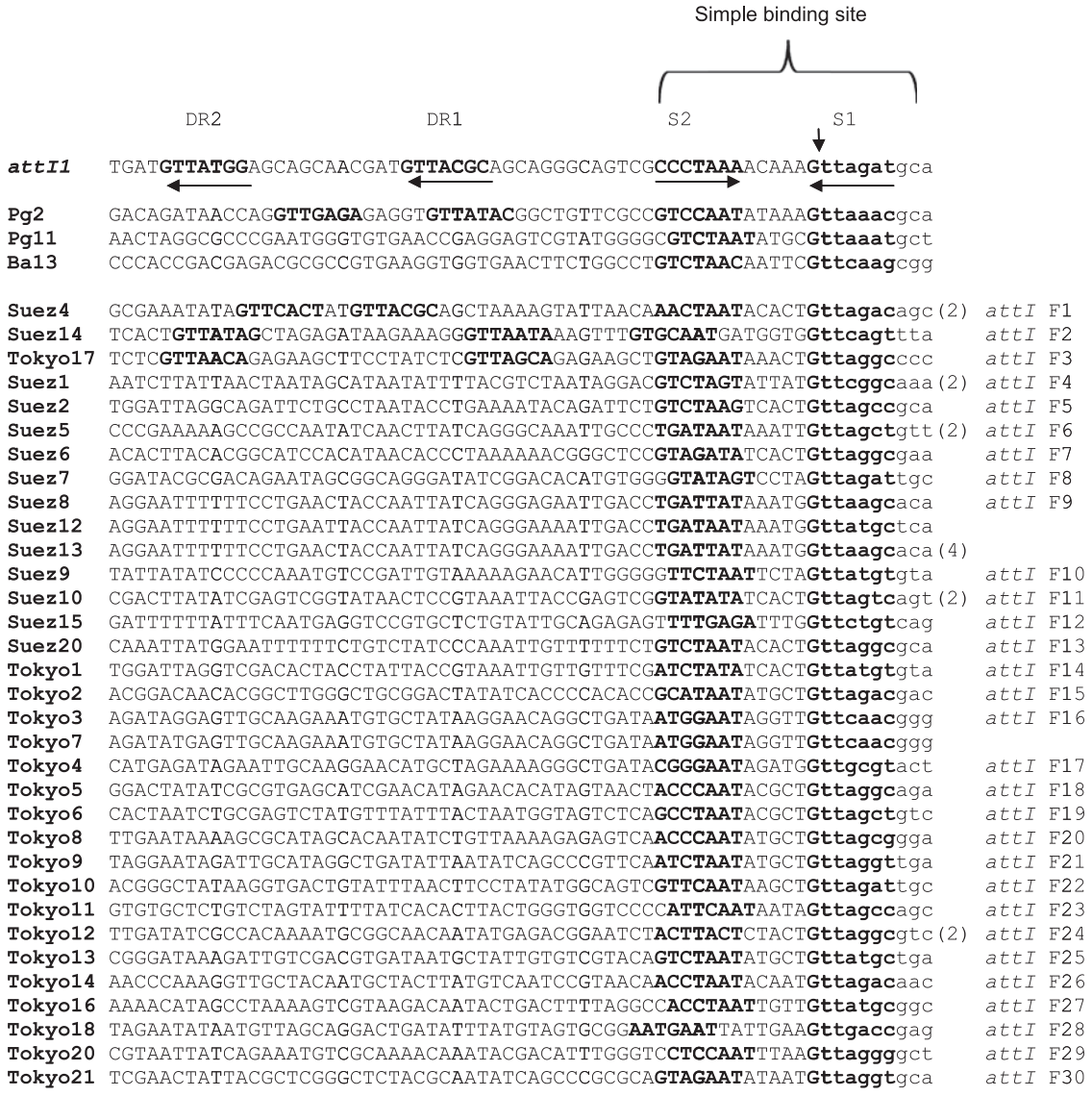


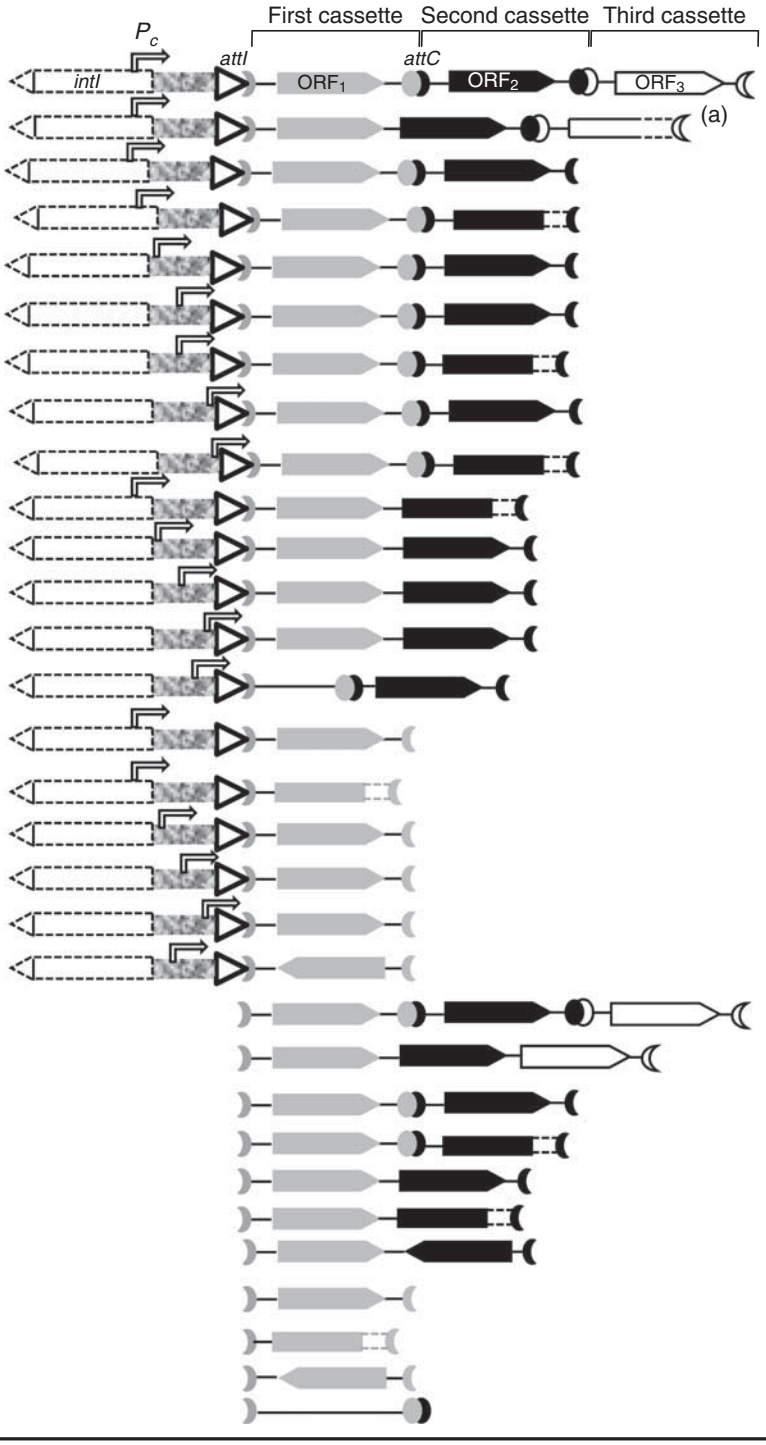
Figure 4 Multiple alignments of recorded putative *attI* families (*attI* F1–F30), which were oriented according to *attI1* length (65 bp) (Partridge *et al.*, 2000) and *attI* sequences, Pg2, Pg11 and Ba13 (Nield *et al.*, 2001). The 7-bp integrase binding sites were in bold, numbered and with orientation indicated by horizontal arrows. A pair of inversely oriented IntI-binding domains, S1 and S2, constituted the putative simple site, while DR1 and DR2 represented core site-like direct repeats. The vertical arrow indicated the known recombination crossover point for insertion of the first gene cassette. The sequence region derived from the first gene cassette was in lower case. Numbers between brackets represented the identical copies of the corresponding *attI* sequence.

the associated direct repeats seen in *attI1*, an arrangement analogous to that seen for *attI3* and in some previously identified *attI* sequences from an environmental DNA (Nield *et al.*, 2001; Collis *et al.*, 2002). Only, three clones, Suez4, Suez14 and Tokyo17, contained putative direct repeat sites, with almost typical sequences, GTTACGC of DR1, for Suez4 and *attI1*; and GTTATA(G)G of DR2, for Suez14 and *attI1* (Figure 4). The absence of direct repeat sequences in most recovered *attI* sites may affect the recombination activity of these sites or suggest that the associated IntI proteins do not rely on them in the same way that IntI1 does (Partridge *et al.*, 2000). However, it has been suggested that these direct repeats act as enhancers by retaining the IntI1 recombinase in the vicinity of the simple site (Gravel *et al.*, 1998).

First recovery of both gene cassette promoter, P_c, and the first gene cassettes associated with marine integrons

Application of current PCR amplification methodology using the primers *intI*-864R and GCP2 achieved two outcomes in terms of integron-associated gene cassette recovery. These were the recovery and identification of the first gene cassette linked to an integron at *attI* and the recovery of the region that normally includes the promoter, P_c, responsible for gene cassette expression. Each of these recovered integron regions were examined for a possible

Table 3 Current integron gene cassette sequence orientations

Orientation no.	Sequence orientation	No. of clones	
		Suez Bay	Tokyo Bay
1		1	
2			1
3			3
4			1
5		1	1
6		7	1
7			2
8			1
9			1
10		2	
11			1
12		1	
13		2	2
14		1	
15		1	4
16		2	
17			2
18		1	1
19		1	
20		1	
21			1
22		1	
23		7	7
24			3
25		9	2
26		2	
27		3	
28		55	5
29		6	
30		2	
31		1	
	Total	107	39

(a) A gene cassette with an incomplete ORF

Table 4 Diversity of gene cassettes within the samples

Gene cassette diversity	*Suez*	*Tokyo*
Total number of cassettes	146 (37+109)[a]	68 (36+32)[a]
Observed richness (cassettes types)	100 (21+79)	63 (32+31)
Chao 1 richness estimator	290.82 (133.86)	491.27 (370.58)
Shannon–Weiner index	3.68 (0.16)	3.39 (0.23)

[a]Number of gene cassettes amplified by primer set *intI*-864R/GCP2+those of amplified by the primer set GCP1/GCP2.

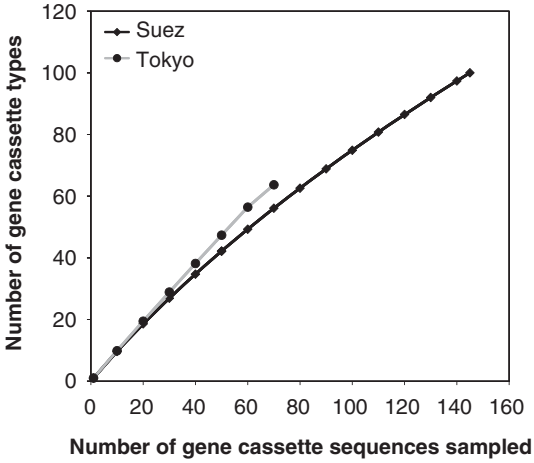


Figure 5 Rarefaction curves showed the richness of gene cassettes at the maximum sample sizes, 146 and 68 gene cassettes, for Suez and Tokyo Bay samples, respectively.

P_c promoter. In all cases, prokaryotic-like promoter sequences with close fit to the consensus TTGACA(17 bp)TATAAT were identified and their relative positions were indicated in Table 3. Most of the integrons found here contained putative P_c promoter sequences, with –35 and –10 regions, having at least three nucleotides of each of the six-base pairs matching to the consensus with typical spacing, a feature similar to that in class 1 integrons (Kim *et al.*, 2007), and a match and spacing likely to allow biological levels of transcription (Zhou and Yang, 2006). Some current P_c promoters had sequences with high similarity to the promoter consensus sequences such as a promoter TTGAGA (17 bp)TAATAG, clone Suez6. Such single-nucleotide polymorphisms may control the strength of P_c promoter and consequently, the level of gene cassette expression (Kim *et al.*, 2007). However, the current P_c promoter sequences and those identified previously in environmental integrons (Nield *et al.*, 2001) have been predicted using software and an experimental evidence is needed to confirm the activity of theses promoters. This study could recover up to three integrated gene cassettes in some cases, including the full structure of the first gene cassette, which was found to vary from 250 to 1254 bp.

How big and diverse are the pools of gene cassettes in Suez and Tokyo Bays?

In addition to the *intI*-864R/GCP2 long PCR that recovered one or more cassettes linked with *intI* genes, another PCR approach, using the primer set GCP1/GCP2, was used, in parallel, for recovering a diverse range of gene cassettes from the two sample sites. Using both PCR approaches, a total of 146 and 68 gene cassettes were recovered and grouped into 100 and 63 cassette types, respectively, representing Suez and Tokyo Bay clone libraries, respectively (Tables 3 and 4). It was concluded, from the Chao 1 richness estimator value, that the total richness of the cassette pool of Tokyo Bay sample was higher than that of Suez (Table 4). This was consistent with rarefaction curves, which implied a trend to more diversity over the same number of sampled cassettes (Figure 5) and with Shannon–Weiner diversity index values (Table 4). Further sampling would be needed to determine if this is a transient phenomenon or one that is stable over time. If the latter, other mobile cassette molecular and ecological analyses would have interest to establish any causal link between these observations.

A total of 17 families of *attC* were detected from current gene cassette surveys (Supplementary Figure 2). Some of the first gene cassettes had complete ORFs but lacked an *attC* site and were followed by a second gene cassette, forming orientation numbers 10, 11, 12 and 13 (Table 3). Usually, lacking of *attC* prevents gene cassettes from being recognized by the integrase enzyme. If these gene cassettes are essential for survival, this may select for the loss of these recombination sites to prevent gene excision. On the other hand, a first gene cassette, SuezGC14.1, was recorded as an empty gene cassette, consisting of only an *attC* without ORF (Table 3, orientation no. 14), an implication of a nonfunctional gene cassette.

Some *attC* sequences from both current sampling sites showed highly similarities and belonged to the same *attC* family (Supplementary Figure 2). Also, six Tokyo Bay gene cassette *attC* sequences, with a same size of 69 bp, showed an average nucleotide identity value of 73.0% with that from a deep-sea hydrothermal vent gene cassette (Elsaied *et al.*, 2007), an observation confirming the existence of conserved *attC* structures in completely different marine habitats.

The gene cassettes encoded proteins with adaptive functions

Of the total 146 and 68 analyzed gene cassettes, 143 and 68 gene cassettes were found to encode predicted proteins from Suez and Tokyo Bay gene cassette pools, respectively. These gene cassettes had predicted ORFs with typical expression orientations (Table 3) as in Stokes *et al.* (2001) and Elsaied *et al.* (2007). That is, the gene was oriented such that expression of P_c is possible. We could identify gene

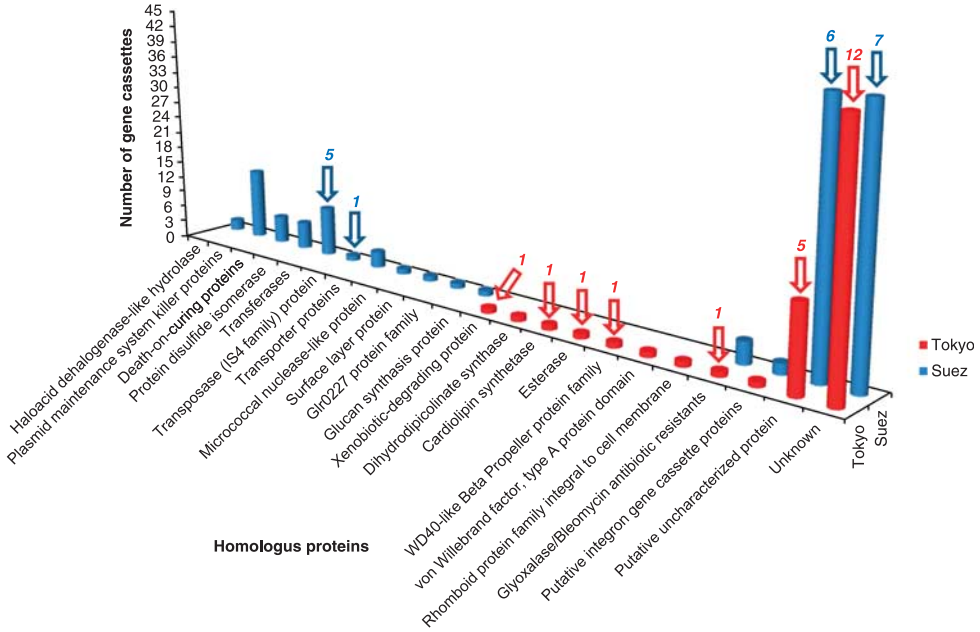


Figure 6 Histogram showed protein coding capacities of gene cassettes recovered from both sampling sites. The down arrows and the labels above the arrows showed the protein coding bars contained first gene cassette proteins and their numbers in each bar, respectively.

cassettes, which encoded proteins with environment adaptive functions (Figure 6). The example for environment adaptive genes, which occurred in both sampling sites, was represented by gene cassettes SuezGC32.1, SuezGC74.1, SuezGC88.1, SuezGC96.1 and the first gene cassette TokyoGC14.1, which encoded proteins with significant homology (*e*-value <0.001) with glyoxalase/bleomycin antibiotic-resistance proteins from *Vibrio parahaemolyticus* and *Bacillus weihenstephanensis*, respectively. The antibiotic bleomycin was first produced by actinomycetes, as an antitumor agent used in clinical settings, and against which there are known resistance genes (Suzuki *et al.*, 1969; Shen *et al.*, 2002). This is the first occurrence of bleomycin resistance genes in marine sediments and it was theoretically expected as both Suez and Tokyo Bay sampling sites are suffering from varieties of anthropogenic pollution. To survive in these niches, each bacterium must develop a resistance to myriad natural antibiotics (D'Costa *et al.*, 2006). Bleomycin resistance genes have been spread to nonclinical environmental bacteria such as those living in activated sludge (Mori *et al.*, 2008). The deduced amino acid identities between bleomycin resistance genes recorded in Suez Bay and that of in Tokyo Bay were low, with an average of 23.0%, representing two different unique families of this gene and supporting the occurrence of these gene variants in nonclinical environments (Mori *et al.*, 2008).

Each sampling site has specific characteristics, implying specific and different niche adaptations. One demonstration of this was in the occurrence of two gene cassettes, SuezGC59.1 and SuezGC93.1, which encoded haloacid dehalogenase-like hydrolases (Figure 6), recording an amino acid identity average of 43.0% with that of *Marinomonas* sp. This enzyme specifically acts on halide bonds in carbon-halide compounds and participates in degradation of both γ-hexachlorocyclohexane and 1,2-dichloroethane, a toxic long-lived organochlorine, which has been detected in the current sampling site of Suez Bay (El-Agroudy *et al.*, 2006). Moreover, occurrence of gene cassettes that encoded haloacid dehalogenases may be correlated with the stress of high concentration of polycyclic aromatic hydrocarbons in Suez sediment (Domínguez-Cuevas *et al.*, 2006; Nemr *et al.*, 2006) (Supplementary Table 1). The detection of five Suez gene cassette-encoded proteins homologous to disulfide isomerase (Figure 6), an enzyme associated with oil sand pollution, supported the role of gene cassette pool as a source for varieties of pollution signal-related genes (Crowe *et al.*, 2001; Koenig *et al.*, 2009). Generally, expression studies on these gene cassettes would be beneficial to assay for the relevant enzyme activities and, consequently, support the current results. However, this study focused on the functional diversity of gene cassettes in coding varieties of adaptive deduced proteins specific for sampling site characteristics.

The Suez gene cassette pool was found to harbor two genomic systems, plasmid maintenance system killer and death-on-curing, which are involved in stabilizing extra-chromosomal DNA (Paul *et al.*, 2005). The plasmid system killer genome was

represented by a total of six clones, each of which carried two gene cassettes, one encoded stable killer protein followed by another gene cassette encoded unstable inhibitor protein, antidote, and gave homology average of 70.0%, with those located in multidrug resistance operons of Gamma- and Delta-proteobacterial species (DeNap and Hergenrother, 2005). One of the five gene cassettes, SuezGC32.2, encoded death-on-curing proteins, was flanked upstream by the gene cassette, SuezGC32.1, which encoded bleomycin antibiotic-resistance protein. It was unclear whether there was a functional relationship between these two adjacent cassettes. However, both the two types of the predicted proteins, bleomycin resistance and death-on-curing, are considered general responses to environmental stresses. The first gene cassette SuezGC15.1 encoded DNA-binding transposase, IS4 protein family, an evidence for a transposon source. Gene cassettes encoded transposases have been recorded in a deep-sea environment, an implication for the role of transposon, as a transfer element, in the mobility and evolution of marine integrons and associated gene cassettes (Elsaied *et al.*, 2007).

The gene cassette TokyoGC20.1 encoded a predicted xenobiotic-degrading protein. Xenobiotics are very often found in the context of the presence of industrial wastes such as dioxins and polychlorinated biphenyls, which commonly occur in Tokyo Bay sediments (Naito and Murata, 2007; Zushi *et al.*, 2010). The location of TokyoGC16.1 as a first gene cassette may facilitate its high expression of the encoded esterase, as responses to toxic metals and antibiotics (Stepanauskas *et al.*, 2005; Brandt *et al.*, 2009). This implication may be supported by the significant homology of the current gene cassette-encoded esterase with that of *Shewanella sediminis*, a marine sediment bacterium that has the ability to degrade several toxic industrial chemical wastes and generates much interest in the field of bioremediation (Zhao *et al.*, 2005). Another first gene cassette, TokyoGC6.1, was found to encode cardiolipin synthetase, an enzyme having stress sensitivity to several drugs and some chemical solvents like toluene (Bernal *et al.*, 2007). The first gene cassette TokyoGC11.1 encoded a unique family of the WD40-like Beta Propeller protein, which is involved in a variety of cell functions such as environment signal transductions (Hudson and Cooley, 2008). Other gene cassettes from both sites encoded a nuclease and a protease rhomboid protein family, key enzymes for digestion of nucleic acids and membrane proteins, respectively, and having a role in the term of biotechnology (Figure 6).

Dominance of first gene cassettes that encoded unknown proteins but predicted signal functions

The current gene cassette pools were dominated by gene cassettes that had no significant homology in databases (Figure 6), a common feature of gene cassettes from environmental metagenomes. Of the total of 42 clones analyzed from both site clone libraries, 11 and 19 first gene cassettes encoded both putative uncharacterized proteins, homologous with those from several bacterial sources, and proteins with no homologs, unknown, respectively (Table 5). Although these gene cassettes included proteins with no identifiable functions, the insertion orientation was such that the corresponding gene had the potential to be expressed from P_c. In the clones Tokyo12 and Tokyo19, a cassette was duplicated, presenting at both the first and second positions. Such a circumstance potentially and actually leads to higher levels of expression of the gene in question compared with the cassette being present only once (Collis and Hall, 1995). The deduced protein encoded by this duplicated gene cassette was predicted as being an unstable protein with a signal functional site, a feature characterizing several unknown gene cassette proteins listed in Table 5. Most of these gene cassette proteins had theoretical isoelectric points $pI > 7$. If the pH of the bacterial surrounding medium is 7, the common medium pH, these protein terminals would have positive charges, which are essential for transfer of signals as an initial step of protein secretion across the membrane (Inouye *et al.*, 1982; Vergunst *et al.*, 2005). These proteins would clearly be interesting targets for further functional characterization.

Biogeographic expansion of gene cassette pool in marine sediments

Two gene cassettes from Suez Bay, SuezGC9.2 and SuezGC38.2 and the Tokyo Bay gene cassette, TokyoGC29.1 showed significant homologies with those recovered from marine sediments in Halifax Harbour, Canada (Koenig *et al.*, 2008). Both Suez and Tokyo Bays, like Halifax Harbour, are urban marine habitats suffering from varieties of anthropogenic pollutions; some of them may be common in the three geographic distant sites. The common occurred pollutant in distant habitats may shape a similar gene cassette structure that can be used as a mobile DNA marker specific for that pollutant. On the other hand, occurrence of glyoxalase/bleomycin resistance gene cassettes in both sites suggested the variable expanding of these gene cassettes in marine sediments. The anthropogenic pollution may accelerate the transfer of these genes from clinical field into natural marine environment. The wide occurrence of gene cassettes that encoded transferase-like enzymes in marine surface sediment from Suez Bay (Figure 6), Halifax Harbour (Koenig *et al.*, 2008), estuaries (Wright *et al.*, 2008) and deep-sea hydrothermal vents (Elsaied *et al.*, 2007) probably suggests the global marine biogeographic distribution of these gene cassette proteins. On the other hand, there was a significant similarity between the gene cassette SuezGC98.2 protein and a methyltransferase from the archaeon *Methanosaeta thermophila*,

Table 5 Characterization of the first GCs that encoded unknown functions

Sample	*GC*	*ORF size (bp)*	*Protein*	*Homology source*	*Protein prediction*			
					Functional sites and protein domain	*Cellular localization*	*Pl*	*Stability*
Suez Bay	SuezGC1.1, 3.1 (99%)[a]	282	Putative uncharacterized protein	*Thauera* sp. (Betaproteobacteria)	No hits reported	Cytoplasmic	5.88	Stable
	SuezGC5.1, 17.1 (98%)	543	Unknown		Signal functional site	Cytoplasmic membrane	9.38	Stable
	SuezGC5.1, 17.1 (98%)	453	Unknown		Signal functional site	Cytoplasmic membrane	6.96	Stable
	SuezGC2.1	696	Putative uncharacterized protein	*Psychrobacter* sp. (Gammaproteobacteria)	No hits reported	Unknown	5.85	Unstable
	SuezGC6.1	252	Unknown		Signal functional site	Unknown	9.52	Stable
	SuezGC7.1	264	Unknown		Signal functional site	Unknown	9.36	Unstable
	SuezGC9.1	246	Putative uncharacterized protein	*Aliivibrio salmonicida* (Gammaproteobacteria)	No hits reported	Unknown	9.61	Stable
	SuezGC10.1	279	Putative uncharacterized protein	*Chlorobium phaeobacteroides* (Chlorobi)	No hits reported	Cytoplasmic	5.64	Stable
	SuezGC12.1	366	Putative uncharacterized protein	*Moritella* sp. (Gammaproteobacteria)	No hits reported	Unknown	6.31	Stable
	SuezGC20.1	396	Unknown		Signal functional site	Cytoplasmic membrane	9.13	Unstable
Tokyo Bay	TokyoGC1.1	318	Unknown		Signal functional site	Unknown	7.70	Stable
	TokyoGC2.1	396	Putative uncharacterized protein	*Chthoniobacter flavus* (Verrucomicrobia)	No hits reported	Unknown	4.96	Stable
	TokyoGC3.1	369	Unknown		No hits reported	Unknown	8.66	Stable
	TokyoGC4.1	303	Putative uncharacterized protein	*Anaerostipes caccae* (Firmicutes)	No hits reported	Unknown	10.45	Unstable
	TokyoGC5.1	321	Unknown		Signal functional site	Cytoplasmic membrane	9.39	Stable
	TokyoGC7.1	468	Unknown		Signal functional site	Unknown	5.73	Unstable
	TokyoGC8.1	606	Unknown		No hits reported	Cytoplasmic	4.77	Unstable
	TokyoGC9.1	510	Putative uncharacterized protein	*Hydrogenivirga* sp. (Aquificae)	Signal functional site	Cytoplasmic membrane	9.81	Stable
	TokyoGC10.1	363	Unknown		Signal functional site	Cytoplasmic membrane	10.09	Unstable
	TokyoGC12.1, 19.1 (98%)	387	Unknown		Signal functional site	Unknown	4.77	Unstable
	TokyoGC13.1	348	Unknown		Signal functional site	Unknown	8.34	Unstable
	TokyoGC15.1	450	Unknown		No hits reported	Unknown	9.27	Unstable
	TokyoGC17.1	747	Putative uncharacterized protein	*Desulfovibrio magneticus* (Deltaproteobacteria)	No hits reported	Unknown	9.10	Unstable
	TokyoGC18.1	234	Unknown		Signal functional site	Unknown	10.21	Unstable
	TokyoGC20.1	720	Unknown		No hits reported	Unknown	10.05	Unstable
	TokyoGC21.1	396	Putative uncharacterized protein	*Planctomyces limnophilus* (Planctomycetes)	No hits reported	Unknown	9.65	Unstable

Abbreviations: GC, gene cassette; ORF, open reading frame.
[a]Amino acid identity average between ORFs of one GC type.

implying an example of lateral gene transfer between bacteria and archaea. These observations may conclude that the biogeographically distribution of gene cassettes in marine environments is affected by several factors, some of which are natural and others are anthropogenic.

This work could improve studying the composition and function of marine integron/gene cassette metagenome. Our current approach is still focusing on addressing several difficulties in studying the marine integrons and associated gene cassettes through improvement of the monitoring and uncovering the ecological role of that metagenome in a wide range of marine environments. In addition, the novelty of the recovered gene cassettes is likely to provide a source of proteins for commercial applications.

Acknowledgements

We are grateful to National Institute of Advanced Industrial Science and Technology, AIST, Japan, and National Institute of Oceanography, Egypt, for offering facilities for marine sediment sampling. We thank Dr Hiroyuki Fuse and the technician Ms Aya Akiba for helping and providing facilities for experiments. The present study has been supported by the Grant-in-Aid for Scientific Research, KAKENHI (no. 19201041) from Japan Society for the Promotion of Science (JSPS), as well as a basic fund from the National Institute of Advanced Industrial Science and Technology, AIST, Japan.

References

Bernal P, Muñoz-Rojas J, Hurtado A, Ramos J, Segura A. (2007). A *Pseudomonas putida* cardiolipin synthesis mutant exhibits increased sensitivity to drugs related to transport functionality. *Environ Microbiol* **9**: 1135–1145.

Berthe-Corti L, Nachtkamp M. (2010). Bacterial communities in-hydrocarbon-contaminated marine coastal environments. In: Timmis KN (ed). *Handbook of Hydrocarbon and Lipid Microbiology*. Springer-Verlag, Berlin, Heidelberg, pp 2350–2357.

Boucher Y, Labbate M, Koenig JE, Stokes HW. (2007). The integron: an adaptable and mobilizable platform that promotes genetic diversity in bacteria. *Trends Microbiol* **15**: 301–309.

Brandt K, Sjøholm O, Krogh K, Halling-Sørensen B, Nybroe O. (2009). Increased pollution-induced bacterial community tolerance to sulfadiazine in soil hotspots amended with artificial root exudates. *Environ Sci Technol* **43**: 2963–2968.

Chao A. (1984). Non-parametric estimation of the number of classes in a population. *Scand J Stat* **11**: 265–270.

Collis CM, Hall RM. (1995). Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* **39**: 155–162.

Collis CM, Kim MJ, Stokes HW, Hall RM. (2002). Integron encoded IntI integrases preferentially recognize the adjacent cognate *attI* site in recombination with a 59-be site. *Mol Microbiol* **46**: 1415–1427.

Crowe A, Han B, Kermode A, Bendell-Young L, Plant A. (2001). Effects of oil sands effluent on cattail and clover: photosynthesis and the level of stress proteins. *Environ Pollut* **113**: 311–322.

D'Costa VM, McGrann KM, Hughes DW, Wright GD. (2006). Sampling the antibiotic resistome. *Science* **311**: 374–377.

DeNap JC, Hergenrother PJ. (2005). Bacterial death comes full circle: targeting plasmid replication in drug-resistant bacteria. *Org Biomol Chem* **3**: 959–966.

Domínguez-Cuevas P, González-Pastor J, Marqués S, Ramos J, de Lorenzo V. (2006). Transcriptional tradeoff between metabolic and stress-response programs in *Pseudomonas putida* KT2440 cells exposed to toluene. *J Biol Chem* **281**: 11981–11991.

El-Agroudy N, El Azim H, Soliman Y, Said T, El Moselhy K. (2006). Concentrations of petroleum hydrocarbons in water and some marine organisms of the Gulf of Suez. *Egyptian J Aquatic Res* **32**: 128–143.

Elsaied H, Maruyama A. (2010). Diversity and role of bacterial integron/gene cassette metagenome in extreme marine environments. In: de Bruijn FJ (ed). *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*. Wiley-Blackwell, Weinheim, Germany.

Elsaied H, Stokes HW, Nakamura T, Kitamura K, Fuse H, Maruyama A. (2007). Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environ Microbiol* **9**: 2298–2312.

Goto M, Kato M, Asaumi M, Shirai K, Venkateswaran K. (1994). TLC/FID method for evaluation of the crude-oil-degrading capability of marine microorganisms. *J Mar Biotechnol* **2**: 45–50.

Gravel A, Fournier B, Roy PH. (1998). DNA complexes obtained with the integron integrase IntI1 at the *attI1* site. *Nucleic Acids Res* **26**: 4347–4355.

Hall RM, Collis CM, Kim MJ, Partridge SR, Recchia GD, Stokes HW. (1999). Mobile gene cassettes and integrons in evolution. *Ann N Y Acad Sci* **870**: 68–80.

Hartmann M, Widmer F. (2006). Community structure analyses are more sensitive to differences in soil bacterial communities than anonymous diversity indices. *Appl Environ Microbiol* **72**: 7804–7812.

Holmes AJ, Gillings MR, Nield BS, Mabbutt BC, Nevalainen KM, Stokes HW. (2003). The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol* **5**: 383–394.

Horii Y, Ohura T, Yamashita N, Kannan K. (2009). Chlorinated polycyclic aromatic hydrocarbons in sediments from industrial areas in Japan and the United States. *Arch Environ Contam Toxicol* **57**: 651–660.

Huang L, Cagnon C, Caumette P, Duran R. (2009). First gene cassettes of integrons as targets in finding adaptive genes in metagenomes. *App Environ Microbiol* **75**: 3823–3825.

Hudson AM, Cooley L. (2008). Phylogenetic, structural and functional relationships between WD- and Kelch-repeat proteins. *Subcell Biochem* **48**: 6–19.

Inouye S, Soberontf X, Franceschini T, Nakamura K, Itakurat K, Inouye M. (1982). Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. *Proc Natl Acad Sci USA* **79**: 3438–3441.

Kim TE, Kwon HJ, Cho SH, Kim S, Lee BK, Yoo HS *et al.* (2007). Molecular differentiation of common

promoters in *Salmonella* class 1 integrons. *J Microbiol Methods* **68**: 453–457.

Koenig JE, Boucher Y, Charlebois RL, Nesbø C, Zhaxybayeva O, Bapteste E *et al.* (2008). Integron-associated gene cassettes in Halifax Harbour, assessment of a mobile gene pool in marine sediments. *Environ Microbiol* **10**: 1024–1038.

Koenig JE, Sharp C, Dlutek M, Curtis B, Joss M, Boucher Y *et al.* (2009). Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney Tar Ponds. *PLoS ONE* **4**: e5276.

Labbate M, Case RJ, Stokes HW. (2009). The integron/gene cassette system: an active player in bacterial adaptation. *Methods Mol Biol* **532**: 103–125.

Lane D, Pace B, Olsen G, Stahl D, Sogin M, Pace N. (1985). Rapid determination of 16S ribosomal sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82**: 6955–6959.

Liu Y, Harrison PM, Kunin V, Gerstein M. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* **5**: R64.

MacDonald D, Demarre G, Bouvier M, Mazel D, Gopaul DN. (2006). Structural basis for broad DNA-specificity in integron recombination. *Nature* **440**: 1157–1162.

Mazel D. (2006). Integrons: agents of bacterial evolution. *Nat Rev Microbiol* **4**: 608–620.

Messier N, Roy PH. (2001). Integron integrases possess a unique additional domain necessary for activity. *J Bacteriol* **183**: 6699–6706.

Mori T, Mizuta S, Suenaga H, Miyazaki K. (2008). Metagenomic screening for bleomycin resistance genes. *Appl Environ Microbiol* **74**: 6803–6805.

Naito W, Murata M. (2007). Evaluation of population-level ecological risks of dioxin-like polychlorinated biphenyl exposure to fish-eating birds in Tokyo Bay and its vicinity. *Integr Environ Assess Manag* **3**: 68–78.

Nemergut DR, Martin AP, Schmidt SK. (2004). Integron diversity in heavy-metal-contaminated mine tailings and inferences about integron evolution. *Appl Environ Microbiol* **70**: 1160–1168.

Nemergut DR, Robeson M, Kysela R, Martin A, Schmidt S, Knight R. (2008). Insights and inferences about integron evolution from genomic data. *BMC Genomics* **9**: 261.

Nemr AE, Khaled A, El-Sikaily A, Said TO, Abd-Allah AM. (2006). Distribution and sources of polycyclic aromatic hydrocarbons in surface sediments of the Suez Gulf. *Environ Monit Assess*; E-pub ahead of print 13 June 2006, doi: 10.1007/s10661-005-9009-4.

Nield BS, Holmes AJ, Gillings MR, Recchia GD, Mabbutt BC, Nevalainen KM *et al.* (2001). Recovery of new integron classes from environmental DNA. *FEMS Microbiol Lett* **195**: 59–65.

Nunes-Duby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A. (1998). Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* **26**: 391–406.

Partridge SR, Recchia GD, Scaramuzzi C, Collis CM, Stokes HW, Hall RM. (2000). Definition of the *attI1* site of class 1 integrons. *Microbiology* **146**: 2855–2864.

Partridge SR, Tsafnat G, Coiera E, Iredell JR. (2009). Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* **33**: 757–784.

Paul D, Pandey G, Jain RK. (2005). Suicidal genetically engineered microorganisms for bioremediation: need and perspectives. *BioEssays* **27**: 563–573.

Robinson A, Guilfoyle A, Harrop S, Boucher Y, Stokes HW, Curmi P *et al.* (2007). A putative house-cleaning enzyme encoded within an integron array: 1.8Å crystal structure defines a new MazG subtype. *Mol Microbiol* **66**: 610–621.

Rodríguez-Minguela CM, Apajalahti JH, Chai B, Cole JR, Tiedje JM. (2009). Worldwide prevalence of class 2 integrases outside the clinical setting is associated with human impact. *Appl Environ Microbiol* **75**: 5100–5110.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Envir Microbiol* **75**: 7537–7541.

Shannon CE, Weaver W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press: Urbana, IL.

Shen B, Du LC, Sanchez C, Edwards DJ, Chen M, Murrell JM. (2002). Cloning and characterization of the bleomycin biosynthetic gene cluster from *Streptomyces verticillus* ATCC15003. *J Nat Prod* **65**: 422–431.

Simberloff D. (1978). Use of rarefaction and related methods. In: Dickson KL *et al* (eds). *Biological Data in Water Pollution Assessment Quantitative and Statistical Analyses*. American Society for Testing and Materials: Philadelphia, pp 150–165.

Singleton D, Furlong M, Rathbun S, Whitman W. (2001). Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Appl Environ Microbiol* **67**: 4374–4376.

Stepanauskas R, Glenn T, Jagoe C, Carytuckfield R, Lindell A, Mcarthur J. (2005). Elevated microbial tolerance to metals and antibiotics in metal-contaminated industrial environments. *Environ Sci Technol* **39**: 3671–3678.

Stokes HW, Hall R. (1989). A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Mol Microbiol* **3**: 1669–1683.

Stokes HW, Holmes AJ, Nield BS, Holley MP, Nevalainen KM, Mabbutt BC *et al.* (2001). Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol* **67**: 5240–5246.

Stokes HW, O'Gorman DB, Recchia GD, Parsekhian M, Hall RM. (1997). Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol Microbiol* **26**: 731–745.

Suzuki H, Nagai K, Yamaki H, Tanaka N, Umezawa H. (1969). On the mechanism of action of bleomycin: scission of DNA strands *in vitro* and *in vivo*. *J Antibiot* **22**: 446–448.

Uchimiya M, Arai M, Masunaga S. (2007). Fingerprinting localized dioxin contamination: Ichihara Anchorage case. *Environ Sci Technol* **41**: 3864–3870.

Vergunst AC, van Lier MC, Dulk-Ras A, Stuve TA, Ouwehand A, Hooykaas PJ. (2005). Positive charge is an important feature of the C-terminal transport signal of the VirB_D4-translocated proteins of Agrobacterium. *Proc Natl Acad Sci USA* **102**: 832–837.

Wright MS, Baker-Austin C, Lindell AH, Stepanauskas R, Stokes HW, McArthur JV. (2008). Influence of industrial contamination on mobile genetic elements: class 1 integron abundance and gene cassette structure in aquatic bacterial communities. *ISME J* **2**: 417–428.

Zhao J, Manno D, Beaulieu C, Paquet L, Hawari J. (2005). *Shewanella sediminis* sp. nov., a novel Na+-requiring and hexahydro-1,3,5-trinitro-1,3,5- triazine-degrading bacterium from marine sediment. *Inter J Syst Evol Microbiol* **55**: 1511–1520.

Zhou D, Yang R. (2006). Global analysis of gene transcription regulation in prokaryotes. *Cell Mol Life Sci* **63**: 2260–2290.

Zushi Y, Tamada M, Kanai Y, Masunaga S. (2010). Time trends of perfluorinated compounds from the sediment core of Tokyo Bay, Japan (1950s–2004). *Environ Pollut* **158**: 756–763.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)