Use of chromosomal integron arrays as a phylogenetic typing system for *Vibrio cholerae* pandemic strains

M. Labbate,¹ Y. Boucher,¹ M. J. Joss,² C. A. Michael,¹ M. R. Gillings² and H. W. Stokes¹

Correspondence

M. Labbate

mlabbate@bio.mq.edu.au

¹Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia

²Department of Biological Sciences, Macquarie University, Sydney, Australia

Approximately 200 serogroups of *Vibrio cholerae* exist, with only two, O1 and O139, responsible for epidemic and pandemic cholera. Strains from these serogroups have evolved from a common progenitor, with lateral gene transfer largely driving their emergence. These strains are so closely related that separation using single- or multi-locus phylogeny has proven difficult. *V. cholerae* strains contain a genetic system called the integron that is located in the chromosome and that can integrate and excise DNA elements called mobile gene cassettes (MGCs) by site-specific recombination. Large arrays of MGCs are found in *V. cholerae* strains. For instance, the O1 El Tor strain N16961 contains 179 MGCs. Since integron arrays are dynamic through recombination and excision of MGCs, it was hypothesized that the MGC composition in a given *V. cholerae* pandemic strain would be useful as a phylogenetic typing system. To address this, a PCR-based method was used to rapidly characterize the MGC composition of *V. cholerae* arrays. The results showed that the MGC composition of pandemic *V. cholerae* cassette arrays is relatively conserved, providing further evidence that these strains have evolved from a common progenitor. Comparison of MGC composition between the *V. cholerae* pandemic strains was also able to resolve the evolution of O139 from a subgroup of O1 El Tor. This level of differentiation of closely related *V. cholerae* isolates was more sensitive than conventional single-gene phylogeny or multi-locus sequence analysis. Using this method, novel MGCs from an O1 classical strain and an Argentinian O139 isolate were also identified, and a major deletion in the MGC array in all pandemic O139 strains and a subset of O1 El Tor strains was identified. Analysis of sequenced *V. cholerae* integron arrays showed that their evolution can proceed by rearrangements and deletions/insertions of large portions of MGCs in addition to the insertion or excision of single MGCs.

Received 8 August 2006

Revised 8 January 2007

Accepted 29 January 2007

INTRODUCTION

Vibrio cholerae is the causative agent of the diarrhoeal disease cholera. Although there are approximately 200 O-antigen serogroups among *V. cholerae* isolates, only two, O1 and O139, are known to cause epidemic and pandemic cholera (Kaper *et al.*, 1995). The O1 serogroup is further classified into two biotypes, namely, the classical and El

Tor biotypes. There have been seven pandemics of *V. cholerae* O1. The sixth pandemic, which ended in 1923, was caused by the classical biotype (Faruque *et al.*, 1998). The El Tor biotype replaced the classical biotype from 1961 onward as the cause of most cholera infections, initiating the beginning of the seventh and current pandemic (Kaper *et al.*, 1995). In 1992, the O139 serogroup emerged by initially displacing the existing O1 El Tor strains as the main cause of cholera in India and Bangladesh (Shimada *et al.*, 1993). However, O1 El Tor strains have re-established their prominence and now share the region with O139 strains.

Laterally acquired genes have largely driven the emergence of new pandemic strains (Faruque & Mekalanos, 2003). The replacement of the O1 classical biotype by the O1 El Tor biotype has been suggested to be due to the acquisition of two chromosomal islands, called VSP-1 and VSP-2,

Abbreviations: 59-be, 59-base element; 6-FAM, 6-carboxyfluorescein; gDNA, genomic DNA; LGT, lateral gene transfer; MGC, mobile gene cassette; VCR, *V. cholerae* repeat sequence.

The GenBank/EMBL/DDBJ accession nos for the sequences reported in this paper are DQ513153–DQ513172 (*recA* sequences), DQ513176–DQ513178 (*V. cholerae* O139 Arg3 gene cassette sequences), DQ513174–DQ513175 (*V. cholerae* O1 classical Z17561 gene cassette sequences) and DQ513173 (sequence of MGC86F/MGC128R PCR product amplified from *V. cholerae* V812).

which probably enhanced epidemic spread (Faruque & Mekalanos, 2003). Furthermore, studies have shown that O139 emerged as a result of lateral gene transfer (LGT) of a fragment of DNA from a serogroup other than O1 into the O-antigen biosynthesis region of O1 El Tor (Ramamurthy *et al.*, 2003). A comprehensive phylogenetic analysis using *mdh* and *groEL* has concluded that consecutive pandemic strains arose from a common O1-serogroup progenitor through the successive acquisition of new virulence regions (O'Shea *et al.*, 2004).

The integron is a 'hotspot' for LGT in the genomes of *V. cholerae* strains (Rowe-Magnus *et al.*, 2003). Integrons possess a gene, *intI*, which encodes a site-specific recombinase (IntI) and an attachment site recognized by this protein, *attI*, into which individual mobile gene cassettes (MGCs) are inserted (Hall *et al.*, 1999). IntI captures MGCs by catalysing recombination between *attI* and an MGC-associated recombination site [59-base element (59-be); also called *attC*, or in *V. cholerae* specifically, the *V. cholerae* repeat sequence (VCR) (Barker & Clark, 1994; Mazel *et al.*, 1998; Stokes *et al.*, 1997)]. Recombination may also occur between two 59-bes. MGCs are the smallest mobilizable elements known, normally only comprising a single gene and the 59-be. In integrons commonly associated with antibiotic resistance, the incorporated MGCs are expressed by an integron promoter (P_c) located adjacent to *attI* in the integron (Collis & Hall, 1995). A similar promoter probably exists in *V. cholerae* integrons, but this has not been demonstrated experimentally. The genome of *V. cholerae* N16961 has been completely sequenced and contains a 126 kb integron that carries 179 MGCs (Rowe-Magnus *et al.*, 2003). Many of these are present in multiple paralogous copies and encode unique products (Rowe-Magnus *et al.*, 2003).

Phylogenetic studies using a single locus or multiple loci have been unable to differentiate between the O1 El Tor and O139 serogroups and between members within these serogroups (Byun *et al.*, 1999; O'Shea *et al.*, 2004; Salim *et al.*, 2005; Stine *et al.*, 2000). In contrast, integron arrays are more dynamic and so may be useful in differentiating closely related strains. The primary goal of this study was to determine whether the integron MGC array would be useful as a phylogenetic typing system for pandemic strains. If the MGC array in *V. cholerae* could be characterized rapidly without having to sequence it in its entirety, it could be an efficient tool to obtain genotypic signatures for different strains. Fingerprinting of strains based on the amplification of MGCs with 59-be primers has recently been utilized for the typing of *V. cholerae* strains in Argentina (Castañeda *et al.*, 2005). Since in that study the PCR products were separated on a standard 2.5% agarose gel, only five to eight bands between 300 and 1000 bp could be resolved per strain, severely limiting the ability to resolve close phylogenetic relationships (Castañeda *et al.*, 2005). Our approach utilized fluorescently labelled 59-be primers and a DNA sequencer with high-resolution denaturing gel electrophoresis for separation and detection

of the amplified PCR products. Approximately 80 bands per strain were resolved and their base length accurately obtained to yield a specific 'fingerprint' based on the MGC size classes present in the strain in question. Comparison of these fingerprints resolved relationships between the O1 El Tor and O139 lineages, and within these serogroups, which previous single- or multi-locus phylogenetic approaches have been unable to do. Furthermore, this method was evaluated as a tool to identify specific MGCs that differ between strains and also strains that carry unique MGCs. Through the use of this method and the analysis of sequenced *V. cholerae* integrons, we found that MGC arrays undergo large rearrangements as well as multi-MGC deletion/insertion events, indicating that MGCs move not only as single units but also as groups of gene cassettes.

METHODS

Bacterial strains. All *V. cholerae* strains used in this study are described in Table 1. Strains were obtained from Dr Antonio Focareta, University of Adelaide, Australia, and stored at -80°C in tryptone soy (TS) broth containing 10% (v/v) glycerol.

Media. *V. cholerae* and *Escherichia coli* cells were routinely grown at 37°C under aerobic conditions in TS broth (Amyl) and Luria-Bertani (LB) broth (Sambrook *et al.*, 1989), respectively.

DNA techniques. Transformation of *E. coli* XL-1 Blue was performed as described elsewhere (Ausubel *et al.*, 1998). Genomic DNA (gDNA) was extracted from overnight cultures using the XS buffer protocol (Tillett & Neilan, 2000). Standard PCR was performed using PCR master mix (Promega) containing 25 U *Taq* DNA polymerase ml^{-1} , 800 μM dNTPs and 1.5 mM MgCl_2 . Primers were used at a final concentration of 0.5 μM each. All primers used in this study are shown in Table 2.

***V. cholerae* MGC size profiling.** Using an adaptation of the method described previously (Michael *et al.*, 2004), the integron arrays of *V. cholerae* strains were profiled according to the presence of MGC size classes. MGCs from the *V. cholerae* arrays were amplified using 6-carboxyfluorescein (6-FAM)-labelled degenerate primers (HS721 and HS722) that anneal to the 59-be sites of MGCs. PCR conditions were one cycle of 80°C for 2 min, 30 cycles of 94°C for 30 s, 50°C for 30 s, 72°C for 90 s, and one cycle of 72°C for 10 min. Samples were run in triplicate and the fluorescently tagged PCR products (containing the MGCs) were analysed on an ABI Prism 377 DNA sequencer to detect the different size classes of MGCs and their relative abundance. The size class of MGCs given here is defined by the length of the PCR product obtained with primers HS721 and HS722. The placement of these primers is such that this length is approximately 82 nt shorter than the actual cassette length as defined by the position of the recombination cross-over point (Stokes *et al.*, 1997). The ABI Prism 377 instrument utilized a standard denaturing sequencing gel with 1200 nt size standards. Electrophoresis was run typically at 25 000 V for ~4 h to resolve products of up to ~1000 nt. Peaks below 70 nt in size were excluded as primers or primer artefacts. Similarly, peaks below a fluorescence value of 30 units were excluded due to their signal-to-noise ratio. Size class data were analysed by generating an output file containing a graphical 'fingerprint' of each PCR size class as a peak separated by size.

Using the known properties of the *V. cholerae* N16961 integron array and its corresponding MGC fingerprint as a reference, the presence and absence of MGC size classes in other *V. cholerae* strains were

Table 1. *V. cholerae* strains used in this study

Strain	Serogroup	Biotype	Serotype	Year of isolation	Source
569B	O1	Classical	Inaba	1946	CVD*
Z17561	O1	Classical	Inaba	1985	ICDDR†
AA14041	O1	Classical	Ogawa	1985	ICDDR
V886	O1	Classical	Ogawa		J. Holmgren‡
N16961	O1	El Tor	Inaba	1971	J. Holmgren
AA14073	O1	El Tor	Ogawa	1985	ICDDR
AA13993	O1	El Tor	Inaba		ICDDR
H-1	O1	El Tor	Ogawa		ICDDR
V62	O1	El Tor	Ogawa		J. Ogg§
C31	O1	El Tor	Ogawa	1957	RIVM
O17SR	O1	El Tor	Ogawa	Pre-1965	CDI¶
C5	O1	El Tor	Ogawa	1957	RIVM
358	O1	El Tor	Inaba		ICDDR
T19479	O1	El Tor	Inaba		J. Holmgren
V812	O1	El Tor	Inaba		NICED#
AI-1837	O139			1993	ICDDR
AJ-18360	O139				ICDDR
232/95	O139				ICDDR
MDO-25	O139			1993	ICDDR
Arg3	O139			1993	CVD

*Center for Vaccine Development, University of Maryland, Baltimore, MD.

†International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh.

‡J. Holmgren, Department of Medical Microbiology and Immunology, University of Goteborg, Sweden.

§J. Ogg, Department of Microbiology, Colorado State University, CO.

||Field isolate belonging to the collection of the Rijksinstituut voor Volksgezondheid en Milieuhygiene (RIVM), Bilthoven, The Netherlands.

¶Central Drug Institute, Lucknow, India.

#Dr S. C. Patel, National Institute of Cholera and Enteric Diseases, Calcutta, India.

determined. The *V. cholerae* N16961 chromosomal integron has been annotated and reported to contain 179 MGCs (Rowe-Magnus *et al.*, 2003); however, our annotation gave 178 MGCs. The gene cassettes of this array are referred to as MGC1–MGC178, with the numbering representing their position in the array beginning with the first cassette located at *attI* (a fully annotated text file is available as Supplementary Table S1). A potential 59-be was identified in MGC157, potentially splitting this cassette into two, thereby making a total of 179 MGCs. However, this 59-be sequence diverges significantly at the 5' end, preventing the binding of the HS722 primer but allowing the binding of HS721, potentially giving an extra PCR product.

In comparisons between *V. cholerae* strains, paired data on peak size (length in nt) and peak height (fluorescence units) were analysed. While relative quantification by comparing differing peak heights within an MGC size-class profile allowed qualitative comparisons to be made, absolute quantification between strains was achieved by applying a correction factor to the compared profiles. This correction factor was obtained by performing quantitative PCR on the single-copy *rpoB* gene [using the *rpoB*1315F and *rpoB*2442R primers (Boucher *et al.*, 2006)] as well as on a selected target-gene cassette present in single copy in N16961. The copy number of the target cassette in other *V. cholerae* genomes was established by examining the quantity cassette: quantity *rpoB* ratio for a strain. Once established that the cassette: *rpoB* ratio was constant (i.e. that the selected cassette was present as a single copy in all genomes), the peak corresponding to the particular cassette was set to an arbitrary value of 100 in all

profiles, and all other peaks within each profile were expressed as a proportion of this peak. The intra- and inter-run reproducibility (i.e. the SD for each peak) was established from the triplicate PCRs for each strain. This allowed statistical-significance levels to be established for comparisons conducted between profiles taken from different strains.

To confirm whether a particular gene cassette was present in a specific *V. cholerae* strain, primers were designed from the sequence of the *V. cholerae* N16961 array. To amplify unique gene cassettes, an internal cassette primer and the corresponding 59-be primer were used. For multi-copy MGCs, an internal cassette primer and primers specific for the cassettes immediately adjacent on either side were used.

Identification of novel gene cassettes in *V. cholerae* strains. To identify MGCs of a unique size class from an individual *V. cholerae* strain, cassettes amplified with HS723 and HS724 (non-labelled versions of HS721 and HS722, respectively) were separated on a long 3% Tris/acetate/EDTA (TAE) agarose gel at 60 V for 6 h. Using a 50 bp ladder as reference, a small agarose block was excised and the nucleic acids extracted from the gel slice using the Qiaquick Gel Extraction kit (Qiagen). The purified nucleic acids were ligated into pGEM-T Easy using the pGEM-T Easy Cloning kit (Promega). The ligation mix was transformed into *E. coli* XL-1 Blue and the clones screened in groups of five by PCR using 6-FAM-labelled primers HS721 and HS722. The amplified PCR fragments were analysed on an ABI Prism 377 DNA sequencer, as described above, for the presence of the correct gene cassette size class. Groups of clones that were positive were subjected to sequencing using the T7 vector primer.

Table 2. Primers used in this study

F, forward; R, reverse.

Primer	Sequence (5'–3')	Target
HS721	AGC CCC TTA RSC GGG CGT TA (6-FAM labelled)	<i>V. cholerae</i> 59-be sites (F)
HS722	TCC CTC TTG ARS CGT TTG TTA (6-FAM labelled)	<i>V. cholerae</i> 59-be sites (R)
HS723	AGC CCC TTA RSC GGG CGT TA	<i>V. cholerae</i> 59-be sites (F)
HS724	TCC CTC TTG ARS CGT TTG TTA	<i>V. cholerae</i> 59-be sites (R)
EpiFOSF	GGA TGT GCT GCA AGG CGA TTA AGT TGG	pCC1FOS sequencing primer (F)
EpiFOSR	CTC GTA TGT TGT GTG GAA TTG TGA GC	pCC1FOS sequencing primer (R)
MGC308bpR	CTT CAC GCG TTG CAG TCG	MGC58, MGC63 and MGC91 primer (R)
MGC57F	GCA GCT AAT GTT GTG TCA AAC G	MGC57 primer (F)
MGC62F	AGA TGC AAG AAA TGA GCA GG	MGC62 primer (F)
MGC90F	ACT ATG GTC ATT CCT TGT TGG	MGC90 primer (F)
MGC5F	TTA GGG CTA ATG TCC GTG TG	MGC5 primer (F)
MGC493bpR	TAA CGT CAC CCG ATT TCT CC	MGC93 and MGC160 primer (R)
MGC92F	CCT GCT CGA TTT GAT ATT GG	MGC92 primer (F)
MGC95R	CTC ATA CAG ATT GAT TGC TGG	MGC95 primer (R)
MGC159F	CGA GCT GAC TTA GCA CAA GC	MGC159 primer (F)
MGC95F	GAC CTC TTG TTT GCA GAG TCC	MGC95 primer (F)
MGC86F	CAG TGG GTT GAT GGT GTT GC	MGC95 primer (F)
MGC128R	CGT CTA CTT CGT GTA ATC GTG C	MGC128 primer (R)
MGC87F	CCA ACT TAA CCT TGA TCG	MGC87 primer (F)
MGC127R	CCA GCC ATG GTG TTG AGG	MGC127 primer (R)
MGC386bpF	AGC CTG ATT TGA ATT GCG	MGC32, MGC113 and MGC121 primer (R)
MGC32/MGC150F	GTT CGC CAG CCT CTT TAC C	MGC32 and MGC150 primer (F)
MGC33R	ATG CCA AAG GAA TCG ACA AC	MGC33 primer (R)
MGC151R	ACC ACA TTG CAC CAC TTA GC	MGC151 primer (R)
MGC110F	TTT ATA ACG AAT TGC CCT CC	MGC110 primer (F)
HS574	GGC GAA GTG GAC GAT ATC GAC C	<i>rpoB</i> in vibrios (F)
HS578	GAT CGA GTC TTC GAA GTT GTA	<i>rpoB</i> in vibrios (R)
<i>rec-1</i>	GAA ACC ATT TCG ACC GGT TC	<i>recA</i> in <i>V. cholerae</i> (F)
<i>rec-2</i>	CCG TTA TAG CTG TAC CAA GCG CCC	<i>recA</i> in <i>V. cholerae</i> (R)
T7	TAA TAC GAC TCA CTA TAG GG	pGEM-T Easy sequencing primer (F)

Fosmid library construction and screening. A fosmid library was constructed from *V. cholerae* Arg3 gDNA using the CopyControl Fosmid Library Production kit (Epicentre). gDNA was extracted using the XS buffer method (Tillett & Neilan, 2000), treated with RNase A (Sigma) and further purified by phenol/chloroform extraction. Fosmid library production relies on gDNA to be sheared into ~40 kb fragments due to the size restriction of the packaging phage capsids. Post gDNA purification, PFGE was undertaken using a 1% agarose gel to ensure that the gDNA was sufficiently sheared to the 40 kb range. Sheared DNA was then end-repaired, ligated into the fosmid vector, packaged into phage capsids and infected into *E. coli* EPI300TM-T1^R, as instructed in the CopyControl Fosmid Library Production kit manual. *E. coli* clones carrying fosmids were plated on LB plates containing 12.5 µg chloramphenicol ml⁻¹. Screening of the fosmid clones was carried out as previously described (Boucher *et al.*, 2006) using HS723 and HS724 primers. Fosmids from positive clones were extracted and the terminal ends of the insert sequenced using the EpiFOSF and EpiFOSR vector primers.

Phylogenetic analysis of *V. cholerae* strains based on *recA* and gene cassette size class composition. To determine whether the MGC size class profile would be useful in resolving relationships between and within *V. cholerae* serogroups, a dendrogram based on MGC size class composition was constructed and compared to a *recA*-based phylogenetic tree. The *recA* gene was amplified from *V. cholerae*

isolates using primers *rec-1* and *rec-2*, as described elsewhere (Stine *et al.*, 2000), and sequenced. Sequences were aligned using CLUSTAL_X (Thompson *et al.*, 1997). Phylogenetic analyses were carried at the DNA level with PAUP* 4.04b, applying the heuristic-search option and using the TBR branch-swapping algorithm. Maximum-likelihood was used as the tree reconstruction method, with the nucleotide substitution model (GTR), gamma rates parameter α , proportion of invariable sites and nucleotide frequencies determined using MODELTEST (Posada & Crandall, 1998). The confidence of each node was determined by building a consensus of 100 maximum-likelihood trees from bootstrap pseudo-replicates of the original data set.

A matrix mapping the absence or presence of a specific MGC size class in all the *V. cholerae* strains was created using data generated by the profiling technique. The matrix was formatted as a two character-state data set, which was used to estimate the relatedness between cassette arrays. This was done in PAUP* 4.04b using a standard distance method (total character differences) as well as parsimony. The confidence of each node was determined by building a consensus tree of 100 distance trees from bootstrap pseudo-replicates of the original data set.

Identification and comparison of *V. cholerae* integron arrays from GenBank. In addition to the completed genome of *V. cholerae* O1 El Tor N16961, there are currently five incomplete *V. cholerae* genome projects in GenBank, which include strains MO10, O395,

RC385, V51 and V52. No serotype information is available for strains RC385, V51 and V52, and only partial gene cassette arrays divided among many small contigs can be identified in their draft genome sequences. Strains MO10 and O395, on the other hand, have been serotyped (O139 and O1 classical, respectively), and the sequences of their entire arrays are available (except for a small gap splitting the O395 array in two contigs). The MO10 and N16961 arrays were sufficiently similar to allow their pairwise alignment using CLUSTAL_W. The O395 array was more divergent and a different approach was used to compare it to the N16961 array. Each cassette identified for O395 was compared against the set of MGCs from N16961 using BLASTN. Cassettes with the best reciprocal BLASTN matches were considered orthologous. Pairwise comparisons between the N16961 cassette array and the arrays of MO10 and O395 were then performed by plotting orthologous cassettes against each other according to their position in the array. In this manner, individual indel and rearrangement events could be inferred.

RESULTS

Reproducibility and evaluation of the MGC size profiling method

The MGC size profiling method was evaluated based on (a) fingerprint reproducibility between independent PCR reactions, and (b) the coverage of MGCs identified using the method against a known standard. With respect to (a), reproducibility between separate PCR reactions was very high, with 85 consistent MGC size classes detected over triplicate samples. The profiles of three independent HS721/HS722 PCR reactions on N16961 gDNA are given to demonstrate reproducibility (Supplementary Fig. S1). With respect to (b), the *V. cholerae* N16961 integron contains 178 MGCs, giving a theoretical size range of 308–3895 bp and a total of 110 potential MGC size classes when amplified with HS721 and HS722. The absence of 25 expected MGC size classes was attributed to (1) non-amplification of some MGCs due to variation in the 59-bp sequences, (2) the inability of the instrument to resolve MGCs above 920 nt, (3) the PCR bias toward amplification of smaller MGCs (300–500 bp), limiting the detection of some larger MGCs, and (4) the shielding of low signal peaks by stronger signal peaks close by. The identification of 85 MGC size classes out of 110 potential MGC size classes gives the method an MGC coverage of ~77% in a given strain.

Validation of the gene cassette size class profiling technique

To confirm that any differences seen between MGC size class profiles reflected strain differences, the profiles of four *V. cholerae* O1 El Tor, Inaba strains were compared. No differences could be observed among strains N16961, T19479 and AA13993. However, in strain V812, at least seven obvious peaks correlating to the sizes of 309, 386, 388, 400, 403, 493 and 496 nt were absent or reduced in peak height compared to N16961 (Fig. 1a). Primers were designed from the *V. cholerae* N16961 integron sequence targeting those MGCs inferred as most likely to correspond

to the above-mentioned peaks. For example, the 309 nt peak was inferred to correspond to the 308 bp MGC size class in N16961, which consists of MGC58, MGC63 and MGC91. Fig. 1(a) shows the MGC size classes and their corresponding MGCs selected for further analysis. PCR amplification of these MGCs was attempted in V812 to determine if they were present or absent. In the case of the 308 bp MGC size class, PCR for MGC91 did not give a product, indicating its absence from V812 and explaining the reduced 309 nt peak height in the V812 profile (the 309 nt peak signal is attributed to amplification of MGC58 and/or MGC63). Using this approach, absent peaks or reduced peak heights in V812 were shown to be due to the absence of MGC91, MGC93, MGC95 and MGC96, and MGC110, MGC113 and MGC121 (see Supplementary Table S2 for a list of MGC numbers and correlating size classes). The close proximity to each other of these missing cassettes suggested that a large deletion in V812 (or a large insertion in N16961) had occurred. Fig. 1(c) plots selected peaks corresponding to MGCs from outside and inside the V812 deletion together with their normalized peak heights. These data clearly indicate a deletion between MGC87 and MGC128. In the profile of V812, peaks corresponding to the single-copy cassettes MGC86 and MGC128 could be identified, indicating that these cassettes were present in V812 and were outside the deleted region. Primers designed for MGC86 (MGC86F) and MGC128 (MGC128R) gave a 2.4 kb product with *V. cholerae* V812 gDNA, confirming that a large indel event of 24528 bp in total had occurred in this region. Sequence analysis of the 2.4 kb PCR product identified that MGC88–MGC126 were missing from the V812 integron array. The 59-bp sequence of MGC87 in V812 was consistent with an integrase-mediated deletion event in N16961. Of the peaks that were found to correspond to a specific MGC size class, all fell within 3 nt of the predicted size, providing a highly accurate size class profile. Small deviations in actual fragment length were attributed to instrument error and A-tailing of PCR products through *Taq* polymerase activity.

The size class profiling technique was therefore able to identify MGCs that were missing when compared to a reference strain such as *V. cholerae* N16961. However, we were also interested in whether the method could identify novel MGCs that were present in other strains but absent in *V. cholerae* N16961. Two unique peaks, with size classes of 374 and 618 nt, were identified in all the O1 classical biotypes tested and were selected for further analysis (Fig. 1b). Sequence information from cassettes in both size classes confirmed that they were not present in the *V. cholerae* N16961 integron array. The 374 nt size class cassette sequence was 99% identical to an MGC sequence (accession no. AF025662) from *V. cholerae* 569B (O1 classical). There was no homologue for the 618 nt size class gene cassette in GenBank.

V. cholerae Arg3 of the O139 serogroup exhibited a distinct size class profile compared to all the other *V. cholerae*

Fig. 1. Comparisons of specific MGC size class peaks between (a) *V. cholerae* O1 El Tor N16961 and O1 El Tor V812, and (b) O1 El Tor N16961 and O1 classical Z17561. MGC size class profiles shown here are representative of three independent runs. Arrows indicate a peak reduced in height or absent in the other strain. The instrument-designated size of the peak (in nt length) is given above the arrow. The table below (a) shows the MGC size class and corresponding gene cassette(s) selected as best candidates for the peaks marked in (a). The MGCs in bold type are absent in V812, as confirmed by PCR. (c) Normalized peak heights of selected MGC size classes between MGC81 and MGC134, easily demonstrating the deletion in V812. The peak heights of V812 are set as a percentage of the N16961 peak heights. *Peak in each data set used to normalize the data from both strains. Black bars, N16961; grey bars, V812.

strains tested, with at least 20 unique size class cassettes compared to N16961 (Fig. 2). A fosmid library of this strain was constructed and seven positive clones were identified carrying gene cassettes. Sequence from the terminal ends of the insert identified novel cassettes. These included a 353 bp size class cassette possessing an ORF with no homologue in the database, a 1461 bp size class cassette including a predicted hypothetical protein with a

trypsin-like serine protease domain, an 852 bp size class cassette carrying two ORFs of which neither had homologues in the database, a 969 bp size class cassette carrying an ORF also with no homologue in the database, and a 374 bp size class cassette found in *V. cholerae* 569B. These data demonstrate the power of the MGC size class profiling technique for rapidly comparing the integron arrays of *V. cholerae* strains.

Fig. 2. Comparison between the MGC size class profiles of (a) *V. cholerae* O1 El Tor N16961, (b) O1 El Tor O17-SR, (c) O139 AJ18360 and (d) O139 Arg3 in the 330–475 nt range. The profile of *V. cholerae* Arg3 was markedly different from that of all other *V. cholerae* O1 and O139 serotypes tested. The MGC size class profiles shown here are representative of three independent runs.

The multi-cassette deletion is found in all O139 strains but only some O1 El Tor strains

Analysis of the gene cassette size class profiles of all 20 *V. cholerae* strains examined revealed that the peaks reduced in height or absent in strain V812 could also be observed in four other O1 El Tor strains (V62, AA14073, H-1 and 358) and in four pandemic O139 strains studied here (AI-1837, AJ-18360, 232/95 and MDO-25). This suggested that these strains also lacked cassettes MGC88–MGC126 in their integron arrays. This was confirmed by PCR with the primers MGC86F and MGC128R, since they generated the same 2.4 kb product seen in *V. cholerae* V812. Except for O139 Arg3 and O1 El Tor C31, both of which displayed markedly different MGC size class profiles, all remaining strains generated a product with primers targeting single-copy MGCs that were part of the deleted fragment in V812 (MGC92F and MGC95R), indicating the presence of cassettes MGC88–MGC126. Since both the O1 classical and a subset of the O1 El Tor biotypes appeared to contain cassettes MGC88–MGC126, it is highly likely that a deletion event occurred in a member of the O1 El Tor lineage rather than two separate insertion events of the same MGCs in members of the O1 El Tor and O1 classical lineages. As a result, this specific event will be referred to as a deletion. The data described here confirm that O139 strains are O1 El Tor derivatives and link O1 El Tor deletion strains more closely to O139 than O1 El Tor strains without the deletion. This specific relationship cannot be resolved by comparison at a single locus, such as the *recA* gene, the sequence of which is identical in all O1 El Tor strains (except for O1 El Tor C31). Even multiple loci are unlikely to describe the specific relationships between these strains, as a study of 26 housekeeping loci from various seventh pandemic clones was not able to differentiate them (Salim *et al.*, 2005). This demonstrates the enhanced resolving power of array characterization in resolving close evolutionary relationships.

The lack of a PCR product with the primer pairs MGC86F/MGC128R and MGC92F/MGC95R in O139 Arg3 and O1 El Tor C31 was not surprising, since the MGC size class profiles of both strains are significantly different from those of the other O1 El Tor and O139 strains, indicating very different MGC arrays. Other studies (Clark *et al.*, 2000) have shown that the integron MGC arrays of C31 and Arg3 are different from those of the other pandemic strains. A deletion in O1 El Tor C31 compared to other O1 *V. cholerae* in a region now identified as the MGC array has been reported in earlier studies (Barker & Clark, 1994; van Dongen & DeGraaf, 1986). Earlier studies have concluded that, unlike O139 Bengal strains, Arg3 is not derived from a *V. cholerae* O1 parent and has evolved independently (Stroeher *et al.*, 1995, 1997). This is supported by the *recA* phylogenetic analysis (Stine *et al.*, 2000; Fig. 3a), and explains the difference in array composition and structure seen here. While C31 is classified as O1 El Tor, the unique MGC size class profile of this strain indicates that it does not form part of the lineage responsible for cholera

pandemics. The *recA* phylogenetic analysis (Fig. 3a) shows that although C31 has diverged from the other O1 El Tor strains, it contains only six mutations in *recA* compared to the O1 El Tor group, suggesting that its MGC array has been quite active and has undergone large changes during its relatively short divergence from the O1 El Tor group. Alternatively, recombination could have occurred in the *recA* gene and/or the region encoding the surface oligosaccharide responsible for its serotype.

Phylogenetic analysis of *V. cholerae* strains based on gene cassette size class composition further resolves the O1 El Tor group

A phylogeny based on the *recA* gene separates pandemic *V. cholerae* strains into three groups: (1) O1 classical, (2) O139 + O1 El Tor and (3) environmental O139 Arg-3 (Fig. 3a). A dendrogram based on the presence and absence of specific MGC size classes (presence/absence matrix is given as Supplementary Table S3) was able to provide finer resolution (Fig. 3b). The MGC deletion divided the O1 El Tor group in two, placing one part closer to the O139 group. This result has more accurately resolved the emergence of O139 from O1 El Tor. As expected, the O1 classical strains were well separated from the O1 El Tor and O139 strains. The O1 classical and O1 El Tor (no deletion) group could be further resolved based on the acquisition or loss of MGCs. As a result, events within the MGC array such as multi-cassette deletions, or the acquisition or loss of MGCs, can be used to map the relationship between closely related *V. cholerae* strains or simply to differentiate them.

Identification of rearrangements and deletions in other *V. cholerae* strains

The presence of a deletion in the MGC array of O1 El Tor V812 and other strains suggested that integron arrays in *V. cholerae* do not necessarily lose or acquire single gene cassettes but are capable of excising a large, linked group of cassettes in a single step. While the stepwise excision of 38 adjacent MGCs cannot be ruled out, it is most likely that a single deletion event is responsible for the absence of MGC88–MGC126 in V812. To examine whether multi-cassette deletions or multi-cassette rearrangements are common, the sequenced MGC arrays of O139 MO10 and O1 classical O395 were further analysed. Pairwise comparison of the O1 El Tor N16961 and O139 MO10 gene cassette arrays identified three major deletion events in the latter (Fig. 4). These events represent losses of two, 15 and 38 cassettes in MO10 corresponding to MGC1–MGC2, MGC40–MGC54 and MGC88–MGC126 (the deletion identified using the MGC size class profiling method) in N16961. The rest of the array is likely to be identical. There are, however, two points at which the comparison cannot be made, in one case due to the gap in the contigs and in another due to the poor sequence quality.

Fig. 3. Evolutionary relationships among the *V. cholerae* strains used in this study based on (a) 726 bp fragment of the *recA* gene (best maximum-likelihood tree with node support values representing the consensus of the best maximum-likelihood trees of 100 bootstrap pseudo-replicates of the data set), and (b) presence/absence of MGC size classes in their integron array [dendrogram representing the most likely relationships between integron arrays based on a two-character state matrix (presence/absence of each MGC size class)]. Line marked as DG indicates the clade containing a common deletion of 38 gene cassettes. The dendrogram presented was constructed by maximum-likelihood, but parsimony methods resulted in similar topology and statistical support.

The array structure of O395 was much less similar to N16961 than the latter was to MO10. Individual events were therefore harder to track in the case of O395. However, some major events were still apparent. A major deletion event appears to have resulted in the loss of 14 contiguous cassettes from O395, including MGC62–MGC75 when compared to N16961 (Fig. 4). Interestingly, multi-MGC rearrangements were also observed, indicating that not only do multiple MGCs excise but such DNA fragments are also capable of reinserting into other locations in the array. This type of event is seen with MGC80–MGC96 in N16961 matching MGC1–MGC11 and MGC13–MGC18 from O395 sequentially (Fig. 4). The absence of MGC12 in O395 is likely to have been due to a later single MGC12 excision event rather than two separate rearrangement events. Other smaller examples of multi-MGC movement can be seen with the cassettes MGC43–MGC45 in N16961 matching MGC52–MGC54 in O395 and MGC1–MGC2 in N16961 matching MGC38–MGC39 in O395 (Fig. 4). These pairwise comparisons strongly suggest that deletion and insertion events can affect groups of MGCs.

DISCUSSION

The main aim of this study was to determine whether MGC composition in *V. cholerae* would be useful in resolving close phylogenetic relationships. MGC composition analysis appears to be particularly useful in typing closely related strains within a species, since MGC arrays are likely to evolve at a faster rate through recombination and excision of MGCs than any given gene through mutation. Use of the *V. cholerae* integron array as a typing system has previously been attempted by amplifying the MGCs in the strain with 59-be primers (called VCR-PCR). However, that study used a standard 2.5 % agarose gel to separate the amplified products, resolving only five to eight bands per strain (Castañeda *et al.*, 2005). This relatively low resolution in

separation and detection of the amplified PCR products limits the ability of the method to resolve close phylogenetic relationships. Our approach utilized 6-FAM-labelled 59-be primers to amplify the MGCs, and a DNA sequencer to separate and detect the fluorescently labelled amplified products. As a result, detection and separation of the amplified products were significantly enhanced, with approximately 80 bands (representing the MGC size classes) identified per strain. Although not all the MGCs can be identified by this method in a given strain, this is offset by the rapidity of the method at providing a quick ‘snapshot’ of the MGC composition of the array. We present this method as a valuable technique for resolving close phylogenetic relationships between *V. cholerae* strains. The method is also useful for isolating novel MGCs or identifying strains carrying novel MGCs.

The MGC size class profiles for O1 classical, O1 El Tor and O139 were relatively conserved (except for O139 Arg3 and O1 El Tor C31), which is consistent with earlier studies (Clark *et al.*, 2000) and confirms that pandemic *V. cholerae* isolates have evolved from a common progenitor. A dendrogram based on the presence/absence of MGC size classes was able to resolve phylogenetic relationships between pandemic *V. cholerae* strains that previous single- or multi-locus phylogenetic studies have been unable to do. PFGE and amplified fragment length polymorphism (AFLP) have provided some differentiation of pandemic strains, although observed differences are not statistically supported and cannot be traced to a specific area of the genome (Arakawa *et al.*, 2000; Kotetishvili *et al.*, 2003; Lan & Reeves, 2002). All differences observed using the fingerprinting method are traceable to the integron array and potentially to specific MGCs. Specifically, the O1 El Tor lineage was able to be subdivided into two groups due to a large 38 MGC deletion in the integron array. The identification of this same deletion in all the O139 strains (except Arg3) linked the deletion O1 El Tor group to O139. This confirmed that O139 is a derivative of O1

Fig. 4. Pairwise comparisons of the *V. cholerae* O1 El Tor N16961 MGC array with (a) O139 MO10 and (b) O1 classical O395. MGCs in the array from compared strains were designated from 1 onwards to indicate their position in the array, and matching MGCs were plotted against one another. Indels and other rearrangements are easily visible and marked according to the key in the figure.

El Tor and maps the emergence of O139 from the O1 El Tor strains carrying the deletion. Further differences in MGC size class composition were able to separate O139 from the deletion O1 El Tor group and strains in the O1 classical and non-deletion O1 El Tor groups. The ability shown here of the MGC array to resolve the close phylogenetic relationships in *V. cholerae* indicates that

the fingerprinting method may be significant in tracing the emergence of future *V. cholerae* pandemic strains.

The identification of a major deletion in the O1 El Tor and O139 lineages was of particular interest as it suggested that MGCs can be excised in groups and not just as single units. Analysis of sequenced MGC arrays from the *V. cholerae* O1 classical, O395 strain and the O139, MO10 strain identified multiple indel events compared to N16961, providing further evidence that MGCs can excise as groups and indicating that this is a relatively common event. In strain V812, where cassettes MGC88–MGC126 have been deleted, analysis of the resulting 59-be between MGC87 and MGC127 is consistent with an integrase-mediated G/TTA excision event. This was also the case for the indel events seen in strains O395 and MO10. While such indel events could be mediated by homologous recombination between two 59-be sites, it is more probable that excision/recombination of multi-MGCs is mediated by the integrase. Furthermore, multi-MGC rearrangements were observed in O395 compared to N16961, suggesting that excision followed by subsequent insertion of multi-MGCs may be relatively common.

While this study has used the MGC size class profiling method for phylogenetic analysis, we have shown that the method can also be utilized to identify both novel MGCs and strains with diverse MGC arrays. When compared to a reference strain such as N16961, deletions and specific MGCs that differ in the tested strain can also be identified. Previous methods that have compared MGC arrays have been unable to do this. As noted above, the VCR-PCR method (Castañeda *et al.*, 2005) does not have the high resolution produced here. A method that utilizes RFLP and a 59-be probe (Clark *et al.*, 2000) is able to identify general conserved fragments between strains; however, specific MGCs differing between strains cannot be identified. When a microarray based on N16961 is used (Dziejman *et al.*, 2005), only comparisons with N16961 can be made, and only MGCs present in N16961 and absent in the tested strain can be identified. Thus, the ability of the method described here to identify the presence or absence of specific gene cassettes has applications in evaluating the MGC pool available to *V. cholerae*. MGCs are a vast source of genetic diversity (Holmes *et al.*, 2003), and a source of new phenotypes, including pathogenicity. For example, a 1099 bp MGC (MGC133 in N16961) in *V. cholerae* encodes a mannose-resistant haemagglutinin that is an important colonization factor (Barker & Clark, 1994; Franzon *et al.*, 1993). Another example of an MGC bearing a pathogenicity determinant is the PAS factor encoded by cassettes found in *Vibrio vulnificus*, *Vibrio alginolyticus* and *Vibrio parahaemolyticus* integrons. The expression of this small protein is preferentially induced during human infection by *V. vulnificus* CMCP98K (Kim *et al.*, 2003), and is believed to mediate secretion of periplasmic proteins required for the *in vivo* survival and pathogenesis of *V. vulnificus* (Lee *et al.*, 2006). Outside the vibrios, acquisition of MGCs has been suggested as being important in the

adaptation and specialization of *Xanthomonas* species into their ecological niches (Gillings *et al.*, 2005). Since LGT has largely driven the emergence of *V. cholerae* pandemic strains, evaluating the size of the MGC pool available to *V. cholerae* may assist in predicting the emergence of future pandemic strains. Assuming that pandemic strains obtain novel MGCs predominantly from non-pathogenic strains in the environment, we anticipate that the MGC size class profiling method will be an excellent tool for evaluating the MGC pool through identification of novel MGC size classes when screening *V. cholerae* isolates.

MGC size class profiling is a rapid method that has uses in resolving phylogenetic relationships between closely related *V. cholerae* pandemic strains that cannot be resolved by examining 16S rDNA or housekeeping genes. Furthermore, the method can be useful to identify novel MGCs and to assess the MGC pool available to *V. cholerae*. Providing that the 59-bp recombination sequences are sufficiently conserved to design primers, this method also has uses in the characterization of MGC arrays in other bacterial species.

ACKNOWLEDGEMENTS

This work was supported by grant 331601 of the National Health and Medical Research Council and by the Macquarie University Centre for Microbial Functional Networks. Y.B. is partially supported by a postdoctoral fellowship of the National Science and Engineering Council of Canada and a Macquarie University Research Fellowship. We are grateful to Dr Antonio Focareta for supplying the *V. cholerae* strains used in this study. Dr Ruiting Lan is thanked for critical reading of the manuscript.

REFERENCES

- Arakawa, E., Murase, T., Matsushita, S., Shimada, T., Yamai, S., Ito, T. & Watanabe, H. (2000). Pulsed-field gel electrophoresis-based molecular comparison of *Vibrio cholerae* O1 isolates from domestic and imported cases of cholera in Japan. *J Clin Microbiol* **38**, 424–426.
- Ausubel, F. A., Brent, R., Kingston, R. F., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1998). *Current Protocols in Molecular Biology*. New York: Wiley.
- Barker, A., Clark, C. A. & Manning, P. A. (1994). Identification of VCR, a repeated sequence associated with a locus encoding a hemagglutinin in *Vibrio cholerae* O1. *J Bacteriol* **176**, 5450–5458.
- Boucher, Y., Nesbo, C., Joss, M., Robinson, A., Mabbutt, B., Gillings, M., Doolittle, W. F. & Stokes, H. W. (2006). Recovery and evolutionary analysis of complete integron gene cassette arrays from *Vibrio*. *BMC Evol Biol* **6**, 3.
- Byun, R., Elbourne, L. D. H., Lan, R. & Reeves, P. R. (1999). Evolutionary relationships of pathogenic clones of *Vibrio cholerae* by sequence analysis of four housekeeping genes. *Infect Immun* **67**, 1116–1124.
- Castañeda, N. C., Pichel, M., Orman, B., Binsztein, N., Roy, P. H. & Centrón, D. (2005). Genetic characterization of *Vibrio cholerae* isolates from Argentina by *V. cholerae* repeated sequences-polymerase chain reaction. *Diagn Microbiol Infect Dis* **53**, 175–183.
- Clark, C. A., Purins, L., Kaewrakon, P., Focareta, T. & Manning, P. A. (2000). The *Vibrio cholerae* O1 chromosomal integron. *Microbiology* **146**, 2605–2612.
- Collis, C. & Hall, R. (1995). Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* **39**, 155–162.
- Dziejman, M., Serruto, D., Tam, V. C., Sturtevant, D., Diraphat, P., Faruque, S. M., Rahman, M. H., Heidelberg, J. F., Decker, J. & other authors (2005). Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system. *Proc Natl Acad Sci U S A* **102**, 3465–3470.
- Faruque, S. M. & Mekalanos, J. J. (2003). Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends Microbiol* **11**, 505–510.
- Faruque, S. M., Albert, M. J. & Mekalanos, J. J. (1998). Epidemiology, genetics, and ecology of toxigenic *Vibrio cholerae*. *Microbiol Mol Biol Rev* **62**, 1301–1314.
- Franzon, V. L., Barker, A. & Manning, P. A. (1993). Nucleotide sequence encoding the mannose-fucose-resistant hemagglutinin of *Vibrio cholerae* O1 and construction of a mutant. *Infect Immun* **61**, 3032–3037.
- Gillings, M. R., Holley, M. P., Stokes, H. W. & Holmes, A. J. (2005). Integrons in *Xanthomonas*: a source of species genome diversity. *Proc Natl Acad Sci U S A* **102**, 4419–4424.
- Hall, R. M., Collis, C. M., Kim, M.-J., Partridge, S. R., Recchia, G. D. & Stokes, H. W. (1999). Mobile gene cassettes and integrons in evolution. *Ann N Y Acad Sci* **870**, 68–80.
- Holmes, A. J., Gillings, M. R., Nield, B. S., Mabbutt, B. C., Nevalainen, K. M. H. & Stokes, H. W. (2003). The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol* **5**, 383–394.
- Kaper, J., Morris, J., Jr & Levine, M. (1995). Cholera. *Clin Microbiol Rev* **8**, 48–86.
- Kim, Y. R., Lee, S. E., Kim, C. M., Kim, S. Y., Shin, E. K., Shin, D. H., Chung, S. S., Choy, H. E., Progulske-Fox, A. & other authors (2003). Characterization and pathogenic significance of *Vibrio vulnificus* antigens preferentially expressed in septicemic patients. *Infect Immun* **71**, 5461–5471.
- Kotetishvili, M., Stine, O. C., Chen, Y., Kreger, A., Sulakvelidze, A., Sozhamannan, S. & Morris, J. G., Jr (2003). Multilocus sequence typing has better discriminatory ability for typing *Vibrio cholerae* than does pulsed-field gel electrophoresis and provides a measure of phylogenetic relatedness. *J Clin Microbiol* **41**, 2191–2196.
- Lan, R. & Reeves, P. R. (2002). Pandemic spread of cholera: genetic diversity and relationships within the seventh pandemic clone of *Vibrio cholerae* determined by amplified fragment length polymorphism. *J Clin Microbiol* **40**, 172–181.
- Lee, J. H., Yang, S.-T., Rho, S.-H., Im, Y. J., Kim, S. Y., Kim, Y. R., Kim, M.-K., Kang, G. B., Kim, J. I. & other authors (2006). Crystal structure and functional studies reveal that PAS factor from *Vibrio vulnificus* is a novel member of the saposin-fold family. *J Mol Biol* **355**, 491–500.
- Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. (1998). A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605–608.
- Michael, C. A., Gillings, M. R., Holmes, A. J., Hughes, L., Andrew, N. R., Holley, M. P. & Stokes, H. W. (2004). Mobile gene cassettes: a fundamental resource for bacterial evolution. *Am Nat* **164**, 1–12.
- O'Shea, Y. A., Reen, F. J., Quirke, A. M. & Boyd, E. F. (2004). Evolutionary genetic analysis of the emergence of epidemic *Vibrio cholerae* isolates on the basis of comparative nucleotide sequence analysis and multilocus virulence gene profiles. *J Clin Microbiol* **42**, 4657–4671.
- Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.
- Ramamurthy, T., Yamasaki, S., Takeda, Y. & Nair, G. B. (2003). *Vibrio cholerae* O139 Bengal: odyssey of a fortuitous variant. *Microbes Infect* **5**, 329–344.

- Rowe-Magnus, D. A., Guerout, A.-M., Biskri, L., Bouige, P. & Mazel, D. (2003). Comparative analysis of superintegrons: engineering extensive genetic diversity in the *Vibrionaceae*. *Genome Res* **13**, 428–442.
- Salim, A., Lan, R. & Reeves, P. R. (2005). *Vibrio cholerae* pathogenic clones. *Emerg Infect Dis* **11**, 1758–1760.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Shimada, T., Balakrish Nair, G., Deb, B. C., Albert, M. J., Sack, R. B. & Takeda, Y. (1993). Outbreak of *Vibrio cholerae* non-O1 in India and Bangladesh. *Lancet* **341**, 1346–1347.
- Stine, O. C., Sozhamannan, S., Gou, Q., Zheng, S., Morris, J. G., Jr & Johnson, J. A. (2000). Phylogeny of *Vibrio cholerae* based on *recA* sequence. *Infect Immun* **68**, 7180–7185.
- Stokes, H. W., O’Gorman, D. B., Recchia, G. D., Parsekhian, M. & Hall, R. M. (1997). Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol Microbiol* **26**, 731–745.
- Stroeher, U. H., Jedani, K. E., Dredge, B. K., Morona, R., Brown, M. H., Karageorgos, L. E., Albert, M. J. & Manning, P. A. (1995). Genetic rearrangements in the *rfb* regions of *Vibrio cholerae* O1 and O139. *Proc Natl Acad Sci U S A* **92**, 10374–10378.
- Stroeher, U. H., Parasivam, G., Dredge, B. K. & Manning, P. A. (1997). Novel *Vibrio cholerae* O139 genes involved in lipopolysaccharide biosynthesis. *J Bacteriol* **179**, 2740–2747.
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.
- Tillett, D. & Neilan, B. A. (2000). Xanthogenate nucleic acid isolation from cultured and environmental cyanobacteria. *J Phycol* **36**, 251–258.
- van Dongen, W. M. A. M. & DeGraaf, F. K. (1986). Molecular cloning of a gene coding for a *Vibrio cholerae* haemagglutinin. *J Gen Microbiol* **132**, 2225–2234.

Edited by: F. A. Rainey