Integrase-directed recovery of functional genes from genomic libraries

Dean A. Rowe-Magnus¹,²,*

¹Division of Clinical Integrative Biology, Sunnybrook Health Sciences Centre, 2075 Bayview Avenue, S1-26A, Toronto, Ontario M4N 3N5 and ²Department of Laboratory Medicine & Pathobiology, Faculty of Medicine, University of Toronto, Toronto, Canada

Received March 9, 2009; Revised June 12, 2009; Accepted June 16, 2009

ABSTRACT

Large population sizes, rapid growth and 3.8 billion years of evolution firmly establish microorganisms as a major source of the planet’s biological and genetic diversity. However, up to 99% of the microorganisms in a given environment cannot be cultured. Culture-independent methods that directly access the genetic potential of an environmental sample can unveil new proteins with diverse functions, but the sequencing of random DNA can generate enormous amounts of extraneous data. Integrons are recombination systems that accumulate open reading frames (gene cassettes), many of which code for functional proteins with enormous adaptive potential. Some integrons harbor hundreds of gene cassettes and evidence suggests that the gene cassette pool may be limitless in size. Accessing this genetic pool has been hampered since sequence-based techniques, such as hybridization or PCR, often recover only partial genes or a small subset of those present in the sample. Here, a three-plasmid genetic strategy for the sequence-independent recovery of gene cassettes from genomic libraries is described and its use by retrieving functional gene cassettes from the chromosomal integron of Vibrio vulnificus ATCC 27562 is demonstrated. By manipulating the natural activity of integrons, we can gain access to the caches of functional genes amassed by these structures.

INTRODUCTION

Integrons are natural gene cloning and expression systems that incorporate circularized open reading frames, called gene cassettes, and convert them to functional genes (1–3). The most notable integron gene cassettes code for antibiotic resistance determinants (4,5) and the stockpiling of these cassettes can produce what are often referred to as resistance integrons (RIs), which contain several antibiotic resistance genes in tandem. Five classes of RIs have been defined based on the homology of the integrase genes (6–12) and each is associated with a mobile DNA element. The paradigm for integron function is the class 1 integron (13). The integron platform consists of an integrase gene and an adjacent primary recombination site (attI). The integrase is a member of the tyrosine recombinase superfamily (14) and catalyses recombination between its cognate attI site and a secondary target called an attC site. The attC site is normally found associated with a single open reading frame (ORF), and the ORF–attC structure is termed a gene cassette (15). Integron gene cassettes are compact structures. They typically consist of the ORF, the attC site and little non-coding sequence (16). Consequently, most gene cassettes are promoterless. However, recombination between the attI and attC sites leads to insertion of the gene cassette downstream of a resident promoter, PC, within the integron that drives expression of the encoded product. The strongest variant of PC constitutes a perfect Escherichia coli consensus promoter and is six times more efficient than the de-repressed P Tac promoter (17). The integron platforms are sedentary, but the common association of integrons with mobile DNA elements such as transposons and conjugative plasmids facilitates the transit of their amassed genes across phylogenetic boundaries, and augments the impact of integrons on bacterial evolution (2,18,19). The proficiency of this partnership is confirmed by the marked differences in codon usage among cassettes within the same RI, indicating that the captured genes are of diverse origins (20–23).

The ancestors of RIs are the chromosomal integrons (CIs) (24,25). The first was discovered in the Vibrio cholerae genome (26). Located on the smaller of the two circular chromosomes, this element spanned 126 kb and gathered at 179 cassettes into a single structure (24,27), dwarfing previously identified RIs. CIs have since been identified throughout the proteobacterial phylum through both systematic searches (25,28–30) and large-scale microbial genome sequencing efforts (27,31–33). CIs can span

*To whom correspondence should be addressed. Tel: +1 416 4806100; Fax: +1 416 4805737; Email: dean.rowe-magnus@sri.utoronto.ca

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
hundreds of kilobases and harbour hundreds of gene cassettes. The CIs of the Vibrionaceae are among the largest identified to date, with those from the clinical \textit{V. vulnificus} strains CMCP6 and YJ016 stretching 153 kb and 138 kb long, and containing 219 and 188 gene cassettes, respectively (34).

Just five \textit{Vibrio} CIs contain a gene cassette reservoir that is equivalent in size to a small genome (35,36). Furthermore, PCR on environmental DNA samples with degenerate primers targeting the conserved regions of \textit{attC} sites (37) has led to the recovery of novel integrase genes and hundreds of diverse gene cassettes (38–40). Importantly, accumulation analysis of the gene cassettes showed no signs of saturation (39), supporting the notion that the integron gene cassette pool is likely immense, and that the process of gene cassette genesis is constant and efficient. Although, the vast majority of the thousands of gene cassettes identified in CIs and environmental samples thus far have no counterparts in the database, or the sole homologues are unassigned ORFs, several have been shown to code for proteins that adopt characteristic three-dimensional structures and assemblies indicative of biological activity (41). The gene cassettes for which an activity has been experimentally demonstrated encoded proteins related to simple adaptive functions, including – metabolism (25,42), virulence (43,44) and antibiotic resistance (45–47). Others appear to be of viral, bacterial or eukaryotic origin, indicating their recruitment from all kingdoms of life (38–40,48). Thus, integron gene cassettes are an important source of new enzymes and novel proteins.

Difficulties in sampling or recognizing integron gene cassettes have limited our capacity to exploit this genetic resource. Although useful, a PCR-based approach for the sequence-independent retrieval of functional gene cassettes from environmental samples is limited by the diversity of the \textit{attC} sites, most of which are unique in sequence and length (25,49). This will undoubtedly preclude the detection of many of the gene cassettes that may be present in a given sample. CI cassettes have been demonstrated to be substrates for recombination by the class 1 integrase enzyme (26,33,45,50,51). Here, the ability of class 1 integrase to recognize diverse \textit{attC} sites is exploited to develop an integron-based genetic selection system to recover functional gene cassettes from genomic libraries.

**MATERIALS AND METHODS**

**Strains, plasmids and media**

The \textit{E. coli} strains, DH10B and EPI300, were from Invitrogen and EpiCentre, respectively. \textit{Vibrio vulnificus} 75.4T (ATCC 27562), \textit{Vibrio cholerae} N16961, \textit{Xanthomonas campestris pv. campestris} 100069T (ATCC 33913) and \textit{Shewanella oneidensis} 106686 (ATCC 700550; MR-1) were obtained from the Pasteur Institute (Paris, France). All strains were grown in LB media (Sigma). The class 1 integron-integrase gene, \textit{intI1}, was amplified with Platinum Pfx polymerase (Invitrogen) using primers \textit{intI1BspHI} (5′-ggaattctgaaacgccgcaactg) and \textit{intI1XbaI} (5′-ctagtctagactctcactagtgaag) and cloned into the NeoI and XbaI sites of pTRC99A. R388 and its \textit{In3} integron have been previously described (5,52,53) and the pCC1FOS fosmid vector was from EpiCentre. Plasmid DNA purifications were performed using a Sigma Plasmid Miniprep and Midiprep kits. Antibiotics were used at the following concentration when indicated: ampicillin (Ap), 100 \mu g/ml; rifampicin (Rf), 100 \mu g/ml; chloramphenicol (Cm), 10 \mu g/ml; kanamycin (Km), 25 \mu g/ml; tetracycline (Tc), 15 \mu g/ml; IPTG 1 mM; L-arabinose (L-ara), 0.02%.

**Overall strategy of the three-plasmid integron selection system**

The overall strategy for the selection system is depicted in Figure 1. The system consists of three plasmids: p1 expresses the class 1 integron-integrase gene, \textit{intI1}; p2 is a self-transmissible recipient plasmid that contains a class 1 integron; p3 is the substrate plasmid into which genomic DNA is cloned and some of the genomic DNA will contain integron gene cassette arrays. The genomic library is constructed in a specialized \textit{E. coli} host strain (βDH10B) that allows for easy antibiotic-independent counter-selection by the simple omission of a specific nutrient from the media. p2 is transferred by conjugation to the \textit{E. coli} strain containing the library. The transconjugants are pooled and transformed with p1 to complete assembly of the system. Integrase-mediated recombination between the \textit{attI1} site of p2 and target \textit{attC} sites on p3 plasmids harbouring integron gene cassettes fuses these plasmids together, and the fusions are selected following conjugation to a second \textit{E. coli} strain and plating on the appropriate selective media that lacks the specific nutrient.

**Constructing the βDH10B host strain**

In \textit{E. coli} strain β2155, the \textit{dapA} gene is disrupted by a construct containing an erythromycin (\textit{erm}) cassette and the \textit{pir} gene (54). The disrupted gene was amplified by PCR with primers DAP1 (5′-atgttcagccggaagtgtg) and DAP2 (5′-cagaaacggcatgtt) and purified. \textit{E. coli} DH10B was transformed with pKOBEGA, which carries the λ red (\textit{redE}\textit{β}) recombination system under the control of the arabinose-inducible \textit{P\textit{BAD}} promoter (55). This system can be used to generate gene knockouts in \textit{E. coli} by allelic exchange. DH10B/pKOBEGA cells were made electrocompetent following arabinose induction and were transformed with the linear DAP1/DAP2 PCR product to create strain βDH10B. Transformants were selected on LB Erm plates containing diaminopimelic acid (DAP). Disruption of the \textit{dapA} gene was confirmed by PCR and the strains inability to grow on plates lacking DAP.

**Fosmid library construction**

Genomic DNA was prepared using the DNAzol genomic DNA isolation reagent (Invitrogen) as per the manufacturer’s instructions. The DNA was then sheared by passage through a Hamilton syringe, end-repaired and ligated into Eco72I-digested pCC1FOS using the CopyControl Fosmid Library Production Kit (EpiCentre). The ligation was then transformed into the βDH10B strain and
colonies were selected on LB DAP Cm plates. Alternatively, sheared DNA was end-repaired and separated on a 1% low melting point PFGE gel. Gel slices containing fragments of 40 kb were excised and digested with gelase (InVitrogen) to recover the DNA. The DNA was concentrated by ethanol precipitation when necessary. The 40-kb DNA fragments were then ligated into Eco72I-digested pCC1FOS. MaxPlax lambda packaging extracts (Epicentre) were used to package the library and transduce βDH10B cells. Transductants were selected on LB DAP Cm plates at 37°C. The colonies were grown overnight at 37°C. To determine how many clones were required in order to obtain a given genome sequence with 99% probability, I used the equation \( N = \ln (1 - P) / \ln (1 - I / GS) \), where \( N \) = number of clones, \( P \) = probability, \( I \) = insert size and GS = genome size (56). Six hundred clones of a 40-kb insert genomic library were assembled. Screening on X-gal indicated that >90% of the clones were recombinant.

**Integrase-mediated recombination of fosmid clones containing gene cassettes**

The R388ISeckKT plasmid was conjugated from E. coli DH5α cells to the βDH10B library. βDH10B transconjugants were selected on LB DAP Cm Km plates at 37°C. These cells were then pooled and transformed with pTRC99A::intI1 and transformants were selected on LB DAP Cm Km Ap plates at 37°C. The specific recovery of fosmid clones containing CI gene cassettes from the library was monitored as follows. The cells were pooled and incubated overnight in LB liquid media containing DAP Cm Km Ap and IPTG at 37°C with shaking. The next day, recombinant plasmids were isolated by conjugation to E. coli EPI300 cells and selection of transconjugants on LB Cm plates that lacked DAP. Higher yield of recombinant plasmids for PFGE, sequencing and activity screening was obtained by growing the transconjugants in media containing the induction solution that increases expression of tefA in the EPI300 strain.

**Sequencing and protein analysis**

Fusion plasmids were used in PCR reactions with primers intI1 and VVR1, VCR1, XCR1 or SPR1-3. The intI1 primer anneals within intI1 and VVR1, VCR1 and XCR1 hybridize to the signature attC sites of V. vulnificus, V. cholerae and X. campestris, respectively (25,28). SPR1, SPR2 and SPR3 hybridize to three of the attC sites from the CI of S. oneidensis (27,52). PCR was conducted as follows: 94°C, 30 s; 60°C, 30 s; 72°C, 2 min, for 30 cycles with Pfu polymerase. Neither primer alone yielded a PCR product, nor did PCR with intI1 and the respective second primer on uninduced cultures. The amplified fragments were sequenced directly with the intI1 primer to confirm integration at the attI site. Sequencing was done at The Centre for Applied Genomics (TCAG, Hospital for Sick Children, Toronto). Homology searches were conducted using BLAST analysis (57), and protein parameters and domain predictions were performed with ProtParam (58) and the PROSITE research tool from the Expert Protein analysis system (59). Protein structure predictions were made using Phyre (60). The 35-kb CI region of V. vulnificus ATCC 27562 was sequenced at The Joint Genome Institute (JGI, University of California) and was deposited at GenBank (accession no. GQ292873).

**RESULTS**

**The three-plasmid integron selection system**

Based on the strategy outlined in Figure 1, a three-plasmid retrieval system was developed for the sequence-independent recovery of complete, functional open reading frames from CIs (Figure 2). The first plasmid, pTRC99A::intI1, expresses the integron-integrase gene, intI1, from the IPTG-inducible P\(_{TRC}\) promoter (61). The second, R388ISeckKT, is the recipient plasmid. It is a Km- and Tc-resistant derivative of the 34-kb self-transmissible plasmid R388, which carries an attI1 site within its class I integron, In3 (52). In3 contains two genes: dfrB2, which confers trimethoprim resistance, and orfA, of unknown function (62). However, dfrB2 is the first gene cassette of In3. Recombination of an attC site with the attI1 site would displace the promoterless dfrB2 from the Pc promoter, potentially leading to trimethoprim sensitivity due to a lack of dfrB2 expression. Furthermore, a defined media that lacks any thymine derivatives must be used when selecting with trimethoprim. The addition of the aph gene permits selection in complex media and avoids the potential problem of clone loss that can occur during growth of genomic libraries in defined media. The plasmid also includes a Tc resistance marker and I-SceI sites, which aid in the isolation of the insert. The third, pCC1FOS, is the substrate plasmid into which genomic DNA is cloned. pCC1FOS is an 8-kb fosmid that carries an F-factor origin of replication (ori2) and partitioning functions that maintain the fosmid in single copy. This circumvents the problem of under-representation of clones carrying genes that may be toxic when present in multiple copies. A second TrfA-dependent origin of replication, oriV, is also present on the fosmid. Genomic fragments of any size can be cloned into the unique Eco72I (blunt end) site within lacZa and recombinants are selected by blue/white screening. If size selection is desired, the fosmid also contains cos sites for packaging into phage particles in vitro, ensuring that the library is comprised only of plasmids containing inserts of ~40 kb under these conditions. The host strain for maintaining libraries is a derivative of the E. coli strain DH10B that has a disruption in dapA, a component of the diaminopimelic acid (DAP) biosynthetic pathway. DAP is incorporated into the bacterial cell wall peptidoglycan. E. coli mutants altered in the DAP pathway require exogenous DAP for growth. Use of this strain (βDH10B) as a donor in conjugation experiments allows for absolute counter-selection in an antibiotic-independent manner even in complex media such as LB, making it easy to isolate transconjugants. βDH10B has a high transformation efficiency for both large and small clones, readily accepts large DNA for the construction of large-insert genomic libraries, and is endA1 and recA1 to ensure high yields of DNA and greater stability of large cloned inserts.
The expression of intI1 catalyzed recombination between the attI1 site of In3 on R388IsceKT and pCC1FOS fosmids from the library that contained inserts with at least one functional attC site, resulting in fusion of the two plasmids. Since the mobilizable R388IsceKT plasmid was KmR, and the CmR pCC1FOS fosmid did not contain an origin of conjugative transfer, the recombinant plasmids were easily isolated by conjugation to a recipient E. coli strain (EPI300-T1R) and selection on LB Km Cm plates. EPI300-T1R cells are endA1 and recA1, and carry a copy-up allele of the trfA gene integrated into the chromosome under control of an inducible promoter. Induced expression of trfA allows for replication from oriV of pCC1FOS. The resulting 40–80-fold increase in the fosmid copy number provided for better fosmid yields for sequencing of the gene cassettes and aided activity-based screening of gene cassettes expressed from P.C of In3. If desired, the integrated fosmid of the recombinant plasmid could be recovered by digestion with I-SceI, self-ligation, transforming EPI300 cells and selecting clones on LB Cm plates. The fosmid harbouring the chromosomal fragment can replicate via the oriV of pCC1FOS.

Recovery of functional genes from a genomic library with the integron-based recombination system

The 153-kb CI of V. vulnificus strain CMCP6 represents 3% of its genome (34). Genomic DNA from strain CMCP6 was ligated into pCC1FOS and the library was either directly transformed into DH10B or packaged into phage and transduced into the host strain. R388IsceKT and pTRC99A::intI1 were then moved to these strains to complete assembly of the three-plasmid system. The expression of intI1 was induced with IPTG and recombinant plasmids were retrieved by conjugation to EPI300-T1R recipient cells. A representative gel of several random recombinant plasmids from the transformed and

---

**Figure 1.** Strategy of the three-plasmid integron selection system. The system consists of three plasmids: p1 expresses the class 1 integron-integrase gene, intI1 (white arrow); p2 is a self-transmissible recipient plasmid that contains a class 1 integron (white arrows); p3 is the substrate plasmid into which genomic DNA is cloned. A genomic library is constructed in p3 (middle panel), some of which will contain integron gene cassettes (p3c; white arrows, orfs; black arrowheads, attC sites). The rest will contain other chromosomal DNA (p3o; black arrows, orfs). A specialized E. coli strain that allows for easy antibiotic-independent counter-selection by the simple omission of a specific nutrient from the media is used as the host strain. p2 is transferred by conjugation (top left panel) to the E. coli strain containing the library. The transconjugants are pooled and transformed with p1 (bottom left panel) to complete assembly of the system (top right panel). Integrase-mediated recombination between the attI1 site of p2 and target attC sites on p3c fuses these plasmids together, while p3o plasmids are not recombined. The fusions are selected following conjugation to a second E. coli strain and plating on the appropriate selective media that lacks the specific nutrient (bottom right panel).
transduced libraries is shown in Figure 3. The wider size range for recombinant plasmids obtained with the transformed library relative to the more uniform plasmid size obtained from the transduced library was reflective of the size-selection imposed by packaging of the ligation reaction into phage particles prior to transduction. Restriction digests of randomly chosen recombinant plasmids suggested that they were indeed the products of recombination between the pCC1FOS library and R388 (data not shown). Specific recombination between the attC site of the substrate plasmid (pCC1FOS library) and the attI1 site of recipient plasmid (R388ISceKT) was verified by PCR and sequencing.

One of the sequenced recombinant clones from the transformed pCC1FOS library contained a 5383-bp insert with 7 attC sites and corresponded to genes \( vv12525-vv12533 \) of strain CMCP6 (Figure 4). One of the partial gene cassettes contained two small overlapping orfs \( vv12525 \) and \( vv12526 \) that coded for a toxin-antitoxin system that we previously characterized from \( V. vulnificus \) strain 27562 (63). Toxin/Antitoxin (TA) systems encode a stable toxin and a specific but unstable antitoxin (64–66). They enhance plasmid maintenance by preventing the proliferation of plasmid-free progeny because the continued de novo synthesis of the short-lived antitoxin is necessary to ensure cell survival. These toxins have proven useful in various molecular biology applications. The TA toxin CcdB (control of cell death), which poisons bacterial DNA gyrase causing degradation of the host chromosome and cell death (67–69), is a key component of the pSW2644T allelic replacement vector, where expression of ccdB is used as a counter-selection marker (70,71). Invitrogen’s Zero Background™ cloning methodology exploits the same toxin to eliminate high background in blunt-end cloning strategies. When an insert is ligated into the vector, the ccdB gene is disrupted, enabling only recombinant colonies to grow.

Another cassette \( vv12532 \) coded for a protein with an Excalibur domain and a putative cold shock domain (CSD) (Figure 4). Excalibur domains bind extracellular calcium and contain the conserved DxDxGxxCE motif, which is strikingly similar to the \( Ca^{2+} \)-binding loop of calmodulin-like EF-hand domains (72). CSD are commonly found in cold shock proteins (CSP) (73).
CSP typically acts as transcription anti-terminators to prevent the formation of secondary structures in RNA molecules, thus facilitating translation at low temperature. Eleven residues that are important for CSP function are proposed to form two RNA-binding motifs, RNP-1 and RNP-2, on the surface of the CSP (74–79). An alignment between VV12532 and the major CSP of E. coli Thermotoga maritima and Bacillus caldolyticus suggested that the RNP-1 and RNP-2 motifs were present in VV12532 (Figure 5). The VV12532 CSD was predicted to adopt an oligomer binding (OB)-fold similar to that of Thermotoga maritima and of the E. coli to adopt an oligomer binding (OB)-fold similar to that of VV12532 (Figure 5). The VV12532 CSD was predicted to form two RNA-binding motifs, RNP-1 and RNP-2, on the surface of the CSP (74–79). An alignment between VV12532 and the major CSP of E. coli, Thermotoga maritima and Bacillus caldolyticus suggested that the RNP-1 and RNP-2 motifs were present in VV12532 (Figure 5). The VV12532 CSD was predicted to adopt an oligomer binding (OB)-fold similar to that of the E. coli, T. maritima and B. caldolyticus CSP, and the surface location of the RNP-1 and RNP-2 motifs was also conserved (Figure 5).

These observations suggested that VV12532 might function as a CSP. At least four separate csp genes must be deleted in E. coli to obtain a cold-sensitive phenotype due to overlapping activity of the CSP family of proteins (80). In E. coli, complementation by any one allele can revert the cold-sensitive phenotype of the quadruple deletion ΔcspABGE strain (81). The genomes of strain YJ016 and CMCP6 contain 4 and 5 CSP homologues, respectively. If a similar situation to that in E. coli occurs in V. vulnificus, targeted disruption of the putative vv12532 csp gene cassette alone would likely have little or no effect on the adaptation of V. vulnificus to cold temperature. Therefore, I opted to test the VV12532 CSP homologue by expressing it in the ΔcspABGE E. coli cold-sensitive mutant. The vv12532 gene was cloned into the arabinose-inducible pBAD24 expression vector (82) and transformed into the E. coli cspABEG quadruple mutant. The cold-sensitive phenotype of the quadruple-deletion strain was suppressed in cells that expresses vv12532 but not in cells that carried the empty vector (Figure 6), suggesting that VV12532 was a CSP. The protein was thus designated CspA\textsuperscript{VVu}. To determine if CspA\textsuperscript{VVu} acted as a transcription anti-terminator, pBAD24::cspA\textsuperscript{VVu} was transformed into the E. coli strain RL211 (83,84). RL211 contains a cat gene that is preceded by the ρ-independent trpL terminator. This strain is sensitive to chloramphenicol but becomes chloramphenicol resistant when transcription termination at trpL is reduced and thus provides a measure of antitermination activity. RL211 cells carrying pBAD24::cspA\textsuperscript{VVu} grew on plates containing chloramphenicol under inducing conditions, whereas the parental strain carrying pBAD24 did not (Figure 6). Hence, CspA\textsuperscript{VVu} was a CSP that appeared to act as a transcription antiterminator.

There were only 27 bp between the end of the 5’ attC site and the start ATG of cspA\textsuperscript{VVu}. A translation initiation region (TIR) was apparent but the sequence did not include a discernable promoter. Hence, I sought to directly screen for the activity of CspA\textsuperscript{VVu} in vivo by relocating gene cassettes from the V. vulnificus CMCP6 CI to the attI1 site of the class 1 integon on R388. CspA\textsuperscript{VVu} was transformed into the E. coli cold-sensitive mutant and transconjugants were selected on LB plates containing Tc and Km at 15°C. Sequence analysis of several transconjugants revealed that a CI cassette, corresponding to cspA\textsuperscript{VVu}, had been precisely inserted into In3.
Comparative genomic analysis of a 35-kb contig containing CI gene cassettes from *V. vulnificus* strain ATCC 27562

Genomic comparisons of the CI cassette contents of the two sequenced *V. vulnificus* strains CMCP6 and YJ016 revealed that they share only 30–39% of their gene cassettes (34,85). The genome of strain 27562 has not been sequenced but we previously demonstrated that it contained a CI (28). To recover fosmids carrying large inserts from the CI, genomic DNA from strain 27562 was ligated into pCC1FOS, the library was packaged into phage and then transduced into the βDH10B host strain. R388IsceKT and pTRC99A::intI1 were then moved to this strain to complete assembly of the system. Recombination was induced with the addition of IPTG and recombinant plasmids were isolated by conjugation to strain EPI300. The 35-kb insert of one of the recovered clones (p2190) was sequenced and annotated (Figure 7), and the *attC* sites and gene cassette boundaries were extracted with the XXR program we developed (28).

Of the 55 complete and two partial gene cassettes that we identified, 39 (71%) contained a discernable *orf* (Table 1), 23 (59%) had homologues in the CIs of other members of the *Vibrionaceae* and only 16 (29%) had homologues in either the CMCP6 or YJ016 CIs. Furthermore, only about 12% of the CI cassettes from

Figure 5. Comparison of the CSD of VV12532 to the major CSP of *E. coli*, *Thermotoga maritima* and *Bacillus caldolyticus*. (Top) Alignment between the CSD of VV12532 and the major CSP of the indicated species. Protein lengths are indicated on the right. Asterisks and colons mark identical and similar amino acids, respectively. The 11 residues that form the two RNA-binding motifs, RNP-1 and RNP-2, and that are important for CSP function are boxed. Bottom, the known structures of the major CSP of *E. coli*, *T. maritima* and *B. caldolyticus* are shown. The two RNA-binding motifs, RNP-1 and RNP-2 (yellow regions), are on the surface of the CSP. A proposed structure for the CSD of VV12532 is also shown.

Figure 6. CspA<sup>vim</sup> is a CSP that causes transcription antitermination. Top, the *E. coli* cspABEG quadruple mutant with pBAD24::vv12532 or the empty vector were streaked on plates containing L-ara and were incubated at the indicated temperatures. Bottom, *E. coli* RL211 cells containing a *cat* gene cassette positioned downstream of the *trpL* terminator were transformed with pBAD24::cspAvvu and streaked on plates containing 25 μg/ml chloramphenicol and arabinose. Plates were incubated overnight at 37°C.

Figure 7. Schematic of a 35-kb region of the *V. vulnificus* ATCC 27562 CI. The insert of p2190 was sequenced and annotated. Gene cassettes with discernable *orfs* (white arrows) are numbered. The *attC* sites are shown as black arrows.
Several of the CI cassettes showed homology to enzymes involved in energy metabolism and detoxification (Table 1). These included a 4-oxalocrotonate tautomerase enzymes involved in energy metabolism and detoxification likely immense.

### Table 1. Homologues of the orfs from the 27562 CI

<table>
<thead>
<tr>
<th>orf</th>
<th>Bacterial species</th>
<th>E-value</th>
<th>Function</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>YP_736195.1</td>
</tr>
<tr>
<td>2</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>YP_001143449.1</td>
</tr>
<tr>
<td>3</td>
<td>Shewanella sp. MR7</td>
<td>8e-30</td>
<td>Preprotein translocase subunit SecA</td>
<td>YP_736195.1</td>
</tr>
<tr>
<td>4</td>
<td>Legionella pneumophila</td>
<td>3e-152</td>
<td>Reverse transcriptase</td>
<td>YP_001251128.1</td>
</tr>
<tr>
<td>5</td>
<td>V. vulnificus</td>
<td>5e-44</td>
<td>Hypothetical protein</td>
<td>NP_93479.1</td>
</tr>
<tr>
<td>6</td>
<td>NH</td>
<td>4e-55</td>
<td>Hypothetical protein</td>
<td>NP_761344.1</td>
</tr>
<tr>
<td>7</td>
<td>Aeromonas salmonicida</td>
<td>6e-31</td>
<td>Hypothetical protein</td>
<td>NP_761251.1</td>
</tr>
<tr>
<td>8</td>
<td>V. vulnificus</td>
<td>2e-17</td>
<td>Hypothetical protein</td>
<td>NP_734249.1</td>
</tr>
<tr>
<td>9</td>
<td>Shewanella sp. MR4</td>
<td>7e-5</td>
<td>Hypothetical protein</td>
<td>NP_001095272.1</td>
</tr>
<tr>
<td>10</td>
<td>V. vulnificus</td>
<td>2e-80</td>
<td>Hypothetical protein</td>
<td>NP_934596.1</td>
</tr>
<tr>
<td>11</td>
<td>V. cholerae</td>
<td>4e-18</td>
<td>Acetyltransferase</td>
<td>YP_002262874</td>
</tr>
<tr>
<td>12</td>
<td>V. parahaemolyticus</td>
<td>3e-108</td>
<td>Esterase/lipase</td>
<td>YP_002156256.1</td>
</tr>
<tr>
<td>13</td>
<td>V. parahaemolyticus</td>
<td>3e-73</td>
<td>Hypothetical protein</td>
<td>NP_761320.1</td>
</tr>
<tr>
<td>14</td>
<td>Algicibrio salmonicida</td>
<td>9e-19</td>
<td>Hypothetical protein</td>
<td>NP_001094978.1</td>
</tr>
<tr>
<td>15</td>
<td>V. parahaemolyticus</td>
<td>9e-131</td>
<td>CBS domain protein</td>
<td>NP_798204.1</td>
</tr>
<tr>
<td>16</td>
<td>V. parahaemolyticus</td>
<td>9e-131</td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>17</td>
<td>V. parahaemolyticus</td>
<td>9e-60</td>
<td>Hypothetical protein</td>
<td>NP_798185.1</td>
</tr>
<tr>
<td>18</td>
<td>Shewanella sp. PV4</td>
<td>5e-59</td>
<td>Permease of the major facilitator superfamily</td>
<td>YP_809940.1</td>
</tr>
<tr>
<td>19</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>20</td>
<td>V. vulnificus</td>
<td>2e-142</td>
<td>Hypothetical protein</td>
<td>NP_761288.1</td>
</tr>
<tr>
<td>21</td>
<td>A. hydrophila</td>
<td>3e-18</td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>22</td>
<td>V. vulnificus</td>
<td>5e-165</td>
<td>Acetyltransferase</td>
<td>NP_809940.1</td>
</tr>
<tr>
<td>23</td>
<td>V. vulnificus</td>
<td>5e-27</td>
<td>Hypothetical protein</td>
<td>NP_132162.1</td>
</tr>
<tr>
<td>24</td>
<td>Photobacterium profundum</td>
<td>2e-17</td>
<td>Hypothetical protein</td>
<td>NP_761288.1</td>
</tr>
<tr>
<td>25</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>26</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>27</td>
<td>V. vulnificus</td>
<td>2e-142</td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>28</td>
<td>A. hydrophila</td>
<td>3e-18</td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>29</td>
<td>NH</td>
<td>4e-47</td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>30</td>
<td>V. vulnificus</td>
<td>5e-27</td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>31</td>
<td>Photobacterium profundum</td>
<td>2e-17</td>
<td>Hypothetical protein</td>
<td>NP_132162.1</td>
</tr>
<tr>
<td>32</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>33</td>
<td>NH</td>
<td></td>
<td>Hypothetical protein</td>
<td>NP_761333.1</td>
</tr>
<tr>
<td>34</td>
<td>V. vulnificus</td>
<td>9e-71</td>
<td>Lactoglutathione lyase</td>
<td>NP_934598.1</td>
</tr>
<tr>
<td>35</td>
<td>V. vulnificus</td>
<td>4e-109</td>
<td>4-oxalocrotonate tautomerase</td>
<td>NP_761336.1</td>
</tr>
<tr>
<td>36</td>
<td>V. vulnificus</td>
<td>1e-86</td>
<td>Adenylate kinase</td>
<td>NP_761336.1</td>
</tr>
<tr>
<td>37</td>
<td>V. vulnificus</td>
<td>7e-81</td>
<td>Adenylate kinase</td>
<td>NP_934722.1</td>
</tr>
</tbody>
</table>

the V. vulnificus CMCP6, YJ016 and 27562 strains had homologues in the CI of V. cholerae N16961. These results support the notion that CIs contribute to the extensive heterogeneity of the Vibrionaceae and that the pool of gene cassettes available for exchange among bacteria is likely immense.

Several of the CI cassettes showed homology to enzymes involved in energy metabolism and detoxification (Table 1). These included a 4-oxalocrotonate tautomerase that enables bacteria to use various aromatic hydrocarbons as their sole energy source (86), a carbohydrate transport protein of the major facilitator superfamily, an adenylate kinase that interconverts ADP to AMP and ATP for driving energy requiring processes and a lactoylglutathione lyase detoxification enzyme that converts toxic methylglyoxal (MG) to lactic acid (87).

To demonstrate that the system could recover integron gene cassettes from different sources within the same sample, genomic DNA from V. cholerae, X. campestris pv campestris and S. oneidensis harbour 216, 23 and three gene cassettes, respectively (27,32,89), and represent 3, 0.3 and 0.1% of the genomes. The three-plasmid system was assembled and integrase-mediated recombination was induced. Recovered fusion plasmids were used in PCR reactions with primers intI1 and VVR1, VCR1, XCR1 or SPR1-3, and the products were sequenced. Products corresponding to integron gene cassettes from the CIs of each species were identified, suggesting that the system can indeed recover low abundance integron gene cassettes from different sources within the same sample.

### Sequencing of integron gene cassette arrays with the integron-based recombination system

The system can also be applied to the sequencing of integron gene cassette arrays. The p2190 plasmid was used as the sole substrate plasmid in the integron-based recombination assay with R388ISceKT. Recombination was induced by the addition of IPTG and recombinant plasmids were isolated by conjugation. The recombinant
Page 9 of 13

Nucleic Acids Research, 2009, Vol. 37, No. 17 e118

plasmids from six randomly chosen transconjugants were digested with either EcoRI or NotI and separated by PFGE (Figure 8). Although each of the plasmids maintained the same overall size, each exhibited a different banding pattern, suggesting that multiple attC sites of p2190 could serve as substrates for recombination with the attI site on R388ISceKT. This can prove useful since a series of plasmids with differing restriction patterns can be sequenced using two sets of unique primers that anneal within the 5' conserved region and the dfrB2 gene of the class 1 integron, and within the cat gene and cos site of pCC1FOS, to obtain the sequence of the gene cassette array (Figure 2).

**DISCUSSION**

Fundamentally, the integron platform is a simple structure. There is an integrase (intI) and an adjacent, cognate recombination site (attI). The substrates for this system are the compact gene cassettes, many of which code for functional proteins or proteins that can adopt a structure of known biological significance. Further interest in the nature of the gene cassette pool has emerged with the demonstration that metabolic (25,42), virulence (43,44) and antibiotic resistance (45–47) genes are structured as gene cassettes integrated at the attI site on R388ISceKT. This can prove useful since a series of plasmids with differing restriction patterns can be sequenced using two sets of unique primers that anneal within the 5' conserved region and the dfrB2 gene of the class 1 integron, and within the cat gene and cos site of pCC1FOS, to obtain the sequence of the gene cassette array (Figure 2).

Here, a genetic system to extract fosmids harboring integron gene cassettes from a bulk genomic pool is described. The system is comprised of three plasmids. The first expresses the class 1 integrase-integrase under the control of an IPTG-inducible promoter. The second is a self-transmissible recipient plasmid that carries an attI site for the integration of gene cassettes via recombination with their attC sites. The third is the donor plasmid that contains the genomic DNA library. IntI1-catalysed recombination between the attI site of recipient plasmid and target attC sites on the inserts of donor plasmids results in the fusion of the two plasmids. The fusion plasmid can then be isolated by conjugation into a recipient host and recovered for sequencing or activity screening of the gene cassettes.

The system has several advantages over existing methods for the detection of gene cassettes. Since integrone-integrases are able to recognize diverse attC sites as targets for recombination (39,51,91,92), a recovery system based on the recombination activity of the integron integrase improves the likelihood of recovering gene cassettes from samples relative to detection by PCR or hybridization. Primers targeting the cos site, dfrB2, cat and intI1 genes can be used to sequence the insert of the fusion plasmid. The I-SceI sites flanking the Km and Tc resistance markers of the fusion plasmid (CmR, KmR, TcR) can be used to isolate clones lacking the R388 backbone (CmR, KmS, TcR) by replica plating on media containing the appropriate antibiotics. Obtaining such clones may be desirable for sequencing purposes or for activity-based screening. Finally, the ability to directly implement activity-based screening is a significant advantage. The activity-based screening of gene banks allows for the recovery of novel enzymes without any prior knowledge of the sequence and relies only on the ingenuity of the screening method. Chitinases, amylases, DNAses, esterase/lipases, proteases, oxygenases, dehydrogenases and antimicrobial activities have been identified via activity screens of genomic libraries (93,94). Many of these enzymes are used as key active components in detergents, various food/feed processes, pulp and paper manufacturing, molecular biology applications and medical treatment protocols. Activity screening requires gene expression and proper folding of the resulting protein in a heterologous host. In E. coli, the minimal set of requirements for gene expression includes the presence of a promoter for transcription and a translation initiation region (TIR) comprised of a Shine–Dalgarno sequence, the initiation codon, and an adequate spacer between them. In such systems, three modes of gene expression can be anticipated: (i) independent gene expression with both the promoter and the TIR provided by the insert; (ii) expression as a transcriptional fusion with only the TIR located on the insert; and (iii) expression as a translational fusion depending on both the promoter and the TIR of the vector. Genes of the last category are virtually inaccessible by shotgun cloning because of the low frequency of functional constructs. Class I integrons naturally incorporate many of the attributes necessary for the proper expression of exogenous genes. The Pc promoter drives expression of promoterless gene cassettes integrated at the attI site. Often, a TIR

---

**Figure 8.** Pulse-field gel of R388-p2190 fusion plasmids. Fusion plasmids that were obtained by integrase-mediated recombination between the integron platform of R388 and the gene cassettes of p2190 were digested with NotI or EcoRI and separated by PFGE. Lanes are as follows: P, PFGE markers; 1 K, 1-Kb markers; 1 and 9, clone 1; 2 and 10, clone 2; 3 and 11, clone 3; 4 and 12, clone 4; 5 and 13, clone 5; 6 and 14, clone 6; 7 and 15, R388; 8 and 16, p2190.
can be recognized at the 5’ boundary of the gene cassette. Even cassettes that have no plausible TIR can be translated. A short open reading frame (ORF-11) that is separated by an octameric spacer from a consensus SD tetramer overlaps the attI site of class I integrons. Hanau-Bercot et al. (95) demonstrated that translation of the TIR-deficient aminoglycoside 6’-N-acetyltransferase (aac(6’)-Ib7) gene cassette was dependent on the translation of ORF-11 but not on the amino acid sequence of the corresponding peptide. These results indicated that ORF-11 and its TIR could substantially enhance expression at the translational level of TIR-deficient gene cassettes. Hence, the integron/gene cassette system can support all three modes of gene expression. Even potential translational attenuation signals, the attC sites themselves (96), are present. Using integron-based activity screening, we previously identified a novel cat gene in the Cl of V. cholera (45) and here I demonstrate the activity of a CSP encoded by a V. vulnificus CI gene cassette.

Greater than 99% of the microorganisms present in many natural environments are unculturable. Hence their genetic wealth has eluded exploitation for medical, biotechnological or basic research purposes. Cultivation-independent approaches, in which DNA is directly isolated from environmental samples and used to construct complex genomic libraries, permits analysis of the genetically and metabolically rich microbial communities in their entirety (97). This obviates the need for cultivation and allows the direct cloning of the collective genomes of all organisms present in a habitat. The quest to identify functional genes in the vast amounts of genetic data contained in these metagenomic libraries presents a variety of challenges, not the least of which is filtering out the often large mounts of undesired DNA that codes for informational (transcription, translation and related processes) and conserved operational (housekeeping) genes (98,99).

This can be overcome, in part, by focusing on genes that are structured as cassettes in integrons, since they tend to code for proteins of unknown function or with tremendous adaptive potential (1,100–104), including structural (105), metabolic (25,42,106), resistance (40,107), toxin (45,46,108) and virulence proteins (44,109). Thus, integron gene cassettes may be a tremendous source of new enzymes and novel proteins from diverse origins. By manipulating the natural activity of integrons, it should be possible to gain access to many of the functional gene cassettes contained within a metagenomic sample.

ACKNOWLEDGEMENTS

I would like to thank Dr Jan-Fang Cheng (JGI) for sequencing of fosmid clones, and Erica Susky and Bella Kushuk for technical assistance.

FUNDING

The Natural Sciences and Engineering Research Council of Canada (293281) and the Canadian Institutes of Health Research (MOP-89794). Funding for open access charge: CIHR.

Conflict of interest statement. None declared.

REFERENCES


