REVIEW ARTICLE

Gene cassettes and cassette arrays in mobile resistance integrons

Sally R. Partridge[1], Guy Tsafnat[2], Enrico Coiera[2] & Jonathan R. Iredell[1]

[1]Centre for Infectious Diseases and Microbiology, University of Sydney, Westmead Hospital, Sydney, NSW, Australia; and [2]Centre for Health Informatics, University of New South Wales, Sydney, NSW, Australia

Correspondence: Sally R. Partridge, CIDM, Level 3, ICPMR Building, Westmead Hospital, NSW 2145, Australia. Tel.: +61 2 9845 6278; fax: +61 2 9891 5317; e-mail: sally.partridge@swahs.health.nsw.gov.au

Received 5 August 2008; revised 17 February 2009; accepted 18 February 2009.
Final version published online 15 April 2009.

DOI:10.1111/j.1574-6976.2009.00175.x

Editor: Eva Top

Keywords
gene cassette; integron; antibiotic resistance; nomenclature; bioinformatics; computer-aided discovery.

Abstract

Gene cassettes are small mobile elements, consisting of little more than a single gene and recombination site, which are captured by larger elements called integrons. Several cassettes may be inserted into the same integron forming a tandem array. The discovery of integrons in the chromosome of many species has led to the identification of thousands of gene cassettes, mostly of unknown function, while integrons associated with transposons and plasmids carry mainly antibiotic resistance genes and constitute an important means of spreading resistance. An updated compilation of gene cassettes found in sequences of such 'mobile resistance integrons' in GenBank was facilitated by a specially developed automated annotation system. At least 130 different (< 98% identical) cassettes that carry known or predicted antibiotic resistance genes were identified, along with many cassettes of unknown function. We list exemplar GenBank accession numbers for each and address some nomenclature issues. Various modifications to cassettes, some of which may be useful in tracking cassette epidemiology, are also described. Despite potential biases in the GenBank dataset, preliminary analysis of cassette distribution suggests interesting differences between cassettes and may provide useful information to direct more systematic studies.

Introduction

Antibiotic resistance in bacteria is a global problem and the genes conferring this resistance have received considerable attention. In Gram-negative bacteria particularly, many of these genes are associated with mobile genetic elements, enabling movement between different DNA molecules (e.g. a bacterial chromosome and a plasmid) and thus transfer between cells, including those of different genera. Genes conferring resistance to many different classes of antibiotics and to disinfectants are found in the form of particular 'genes cassettes' that collectively form an important gene pool. These cassettes can exist transiently in a free circular form (Collis & Hall, 1992) but do not include all of the functions necessary for their own movement and are usually associated with gene capture and expression elements called integrons (Stokes & Hall, 1989).

A gene cassette typically consists of little more than a single promoter-less gene and a recombination site. These recombination sites differ in length and sequence but share conserved regions at their ends and are generally imperfect inverted repeats predicted to form stem–loop structures. They were initially termed 59-base elements from a consensus of the first examples to be identified (Cameron *et al.*, 1986) and this name was retained even when it became apparent that their lengths are quite variable (Hall *et al.*, 1991). The term *attC* site (Hansson *et al.*, 1997), consistent with terminology used in other site-specific recombination systems, has since been widely adopted and will be used in the remainder of this review.

An integron is generally defined by the presence of an *intI* gene, encoding an integrase (IntI) of the tyrosine recombinase family, and an *attI* recombination site. IntI-catalysed recombination between *attI* and/or *attC* sites results in insertion or excision of cassettes (Fig. 1). Several cassettes may be inserted in tandem in the same integron to create an array, with cassettes always inserted in the same orientation. Integrons were first identified as a result of their association with antibiotic resistance genes and mobile elements (Stokes & Hall, 1989), but it is now clear that they are found on the chromosome of many species and constitute an important and near-ubiquitous class of genetic elements (Mazel, 2006; Boucher *et al.*, 2007).

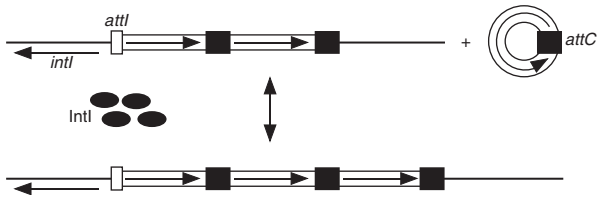

Fig. 1. Integration and excision of gene cassettes by site-specific recombination. IntI encoded by the *intI* gene in the integron catalyses recombination between the *attI1* site (open box) of the integron and/or the *attC* site(s) of gene cassette(s) (black box) resulting in insertion or excision of a cassette. Horizontal arrows indicate the opposite orientations of *intI* and cassette-borne genes.

The amino acid sequences of IntI integrases have been used as a basis for dividing integrons into 'classes', with those carrying *intI1* defined as 'class 1', *intI2* as 'class 2', *intI3* as 'class 3', etc. *intI1*, *intI2* and *intI3* were first identified in association with mobile genetic elements and *intI4* and others with chromosomal integrons. However, *intI1* (Stokes *et al.*, 2006; Gillings *et al.*, 2008) and *intI3* (Xu *et al.*, 2007) have recently been found in different contexts in environmental isolates and a 'chromosomal' *intI* appears to have been transferred to a *Vibrio cholerae* plasmid (Szekeres *et al.*, 2007). Thus the same *intI* gene may be found as part of different structures, and using the class designation alone, which does not indicate the context of *intI*, may no longer be sufficient. Here, we use the term 'mobile resistance integrons' (MRI) to refer to those with *intI1*, *intI2* or *intI3* that carry mainly antibiotic resistance genes and are associated with mobile or potentially mobile elements and we largely restrict our analysis to cassettes found in these integrons.

Chromosomal integrons typically have long arrays with related *attC* sites that exhibit some species specificity (Rowe-Magnus *et al.*, 2001). For example, the *attC* sites of the *V. cholerae* chromosomal integron were initially identified as *V. cholerae* repetitive DNA sequences (VCR) (Barker *et al.*, 1994). In contrast, MRI generally contain few cassettes but the associated *attC* sites may be quite varied. Some *attC* sites found in cassettes carried by MRI are closely related to those in chromosomal integrons, for example 'group 1' (Recchia & Hall, 1997) or 'classical' *attC* sites are related to those found in *Xanthomonas* chromosomal integrons (Rowe-Magnus *et al.*, 2001). Thus, although the mechanism by which cassettes are assembled remains unknown, this process may take place in different species with chromosomal integrons, which then act as reservoirs of cassettes that can be acquired by MRI (Rowe-Magnus *et al.*, 2001). Completely unrelated genes may be associated with similar *attC* sites, probably reflecting acquisition from a common chromosomal integron or species (Rowe-Magnus *et al.*, 2001). Closely related genes may also be associated with almost identical *attC* sites, suggesting divergence from a common ancestral cassette, or with different *attC* sites, suggesting different origins (Recchia & Hall, 1997).

Capture of cassettes and expression of cassette-borne genes

IntI-mediated site-specific recombination has important differences from reactions mediated by other tyrosine recombinases. A typical tyrosine recombinase 'simple site' is minimally comprised of a pair of highly conserved 9–13-bp inverted integrase binding sites separated by a 6–8-bp spacer region. Two such sites are recombined by sequential strand exchange events at the 5′ boundaries of the spacer and identity between the spacers of the recombination partners is generally required for activity (Grindley *et al.*, 2006). In contrast, the most efficient reaction catalysed by an IntI integrase (Collis *et al.*, 1993, 2001) appears to be that between two architecturally distinct sites, the cognate *attI* site and the *attC* site of a gene cassette, and only a single strand exchange occurs. Each *attI* site includes one simple site and two additional integrase binding sites (Collis *et al.*, 1998; Gravel *et al.*, 1998; Partridge *et al.*, 2000), while the *attC* sites are more complex and variable.

attC sites structure and the recombination reaction

An *attC* site contains two simple sites, each composed of a pair of conserved 'core sites' (7 or 8 bp), referred to as 1L and 2L, 2R and 1R (Stokes *et al.*, 1997) or R″ and L″, L′ and R′ (Recchia & Sherratt, 2002). 1L and 2L are separated by a 7-bp spacer and 2R and 1R by a 7- or 8-bp spacer (Fig. 2a). The 1L/2L and 2R/1R pairs are separated by a central region that varies in length and sequence between different *attC* sites. 1L and 1R are usually reverse complements of one another and 2L and 2R are generally complementary except for an extra base present in 2L. The spacer regions are not complementary, but the central region usually forms an imperfect inverted repeat. The GTT of 1R (and complementary AAC of 1L) are completely conserved and the recombination site is between the G and first T of 1R (dotted line in Fig. 2a). In the linear integrated form of a cassette opened up at this position, the G of 1R defines the end of the cassette and remainder of 1R is found at the start of the cassette, separated from the rest of the *attC* site by the cassette gene.

The current model for IntI1-catalysed site-specific recombination, summarized in Mazel (2006), involves the bottom strand of the *attC* site only (Francia *et al.*, 1999), folded into a bulged hairpin structure (Fig. 2b) (Bouvier *et al.*, 2005; MacDonald *et al.*, 2006). The folded *attC* recombines with a double-stranded *attI1* site or another folded *attC* site and a subsequent replication step is required to resolve the Holliday junction intermediate (Bouvier *et al.*, 2005; MacDonald *et al.*, 2006).

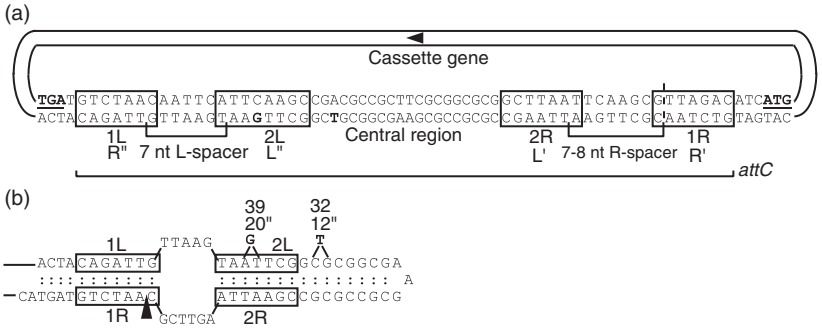

Fig. 2. *attC* site architecture. (a) Double-stranded circular form of the *aadA7* cassette showing the sequence of the *attC* site. The direction of the *aadA7* gene is indicated by a horizontal arrowhead and the start and stop codons on the top strand are in bold and underlined. Core sites are boxed and labeled, with the end points of the spacers indicated and extrahelical bases on the bottom strand are in bold. The position at which the cassette 'opens up' to give the linear form is shown by a vertical dotted line. (b) Folded bottom strand of the *aadA7 attC* site showing the extrahelical bases 'flipped out'. The position of the recombination crossover is shown by a vertical arrowhead.

Conservation of the first three bases (GTT) of the 1R site and the complementary AAC of 1L is important for *attC* site activity, but the identity of the remaining bases in these sites is less critical (Stokes *et al*., 1997). Sequence complementarity in and flanking the core sites (Biskri *et al*., 2005) appears very important, as are two extrahelical bases that are 'flipped out' of the folded structure (Johansson *et al*., 2004; MacDonald *et al*., 2006) (Fig. 2b). One, labeled 39 in Johansson *et al*. (2004) and 20″ in Demarre *et al*. (2007), corresponds to the extra base in 2L compared with 2R. While this is G in the *attC* sites examined in detail, base substitution here does not appear to greatly affect integrase binding (Johansson *et al*., 2004). The second extrahelical base, at position 32 (or 12″), is needed for optimal integrase binding but its positioning may be less important than that of G39 (Johansson *et al*., 2004). The five bases of 2L and 2R furthest from the central spacer region appear to be important for function, but mismatches in the remainder of these sites seemed to have less effect (Johansson *et al*., 2004).

Expression of cassette-borne genes

Most cassettes include only a short region between the end of the 1R site and the predicted start codon of the cassette gene. A suitably spaced ribosome-binding site (RBS) can usually be identified in this region, but only a few cassettes, for example *qac* (Guerineau *et al*., 1990), *cmlA* (Bissonnette *et al*., 1991), and *ereA* (Biskri & Mazel, 2003), appear to carry a promoter. Expression of most cassette-borne genes thus relies on their location in an integron. In class 1 integrons, a promoter (Pc; formerly Pant) located within *intI1* drives expression of cassette genes (Collis & Hall, 1995). In some cases, a second promoter (P2) is created by insertion of three G residues that increase the spacing between potential -35 and -10 sites to the optimum 17 bp (Collis & Hall, 1995). Several Pc variants have been identified and shown to result in different levels of expression (Lévesque *et al*., 1994; Bunny *et al*., 1995; Collis & Hall, 1995; Brízio *et al*., 2006; Papagiannitsis *et al*., 2008). A similar promoter has been demonstrated in a class 3 integron (Collis *et al*., 2002).

Analysis of transcripts originating from Pc suggests that the stem–loop structures formed by *attC* sites might be acting as transcription terminators, so that the position of a cassette in an array may be an important determinant of cassette gene expression (Collis & Hall, 1995). However, little is currently known about whether *attC* sites of different lengths and sequences have noticeably different effects on the expression of downstream genes. Recombination between *attC* and *attI* appears to be the preferred reaction, resulting in insertion of an incoming cassette at the start of an array, closest to Pc. Cassettes have also been observed to 'move up' to the first position in an array following exposure to the relevant antibiotic (e.g. Rowe-Magnus *et al*., 2002). Such rearrangement of cassettes in an array could potentially occur via a number of IntI-mediated processes or by homologous recombination (Hall & Collis, 1995).

Structure of MRI

Integrons with *intI1*

The ancestor of mobile class 1 integrons may have been generated by acquisition of *intI1* and *attI1* by a transposon of the Tn*5053* family (Stokes *et al*., 2006; Gillings *et al*., 2008) to give a structure related to Tn*402* (also called Tn*5090*; Radstrom *et al*., 1994). Tn*402* contains a complete *tni* transposition region (Fig. 3a), making it both a functional transposon and an integron, and is bounded by 25-bp inverted repeats (IRi, integrase end; IRt, *tni* end). The specific sequence from IRi to the start of the first cassette is referred to as the 5′-conserved segment (5′-CS).

The most frequently identified type of class 1 integrons retain the 5′-CS but include part of a region referred to as

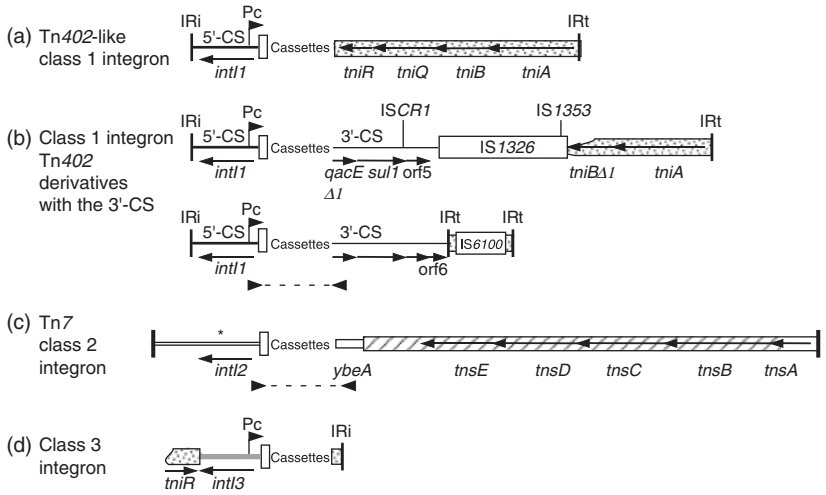

Fig. 3. Structures of MRI. Filled vertical bars present inverted repeats, *attI* sites are shown as small open boxes (not to scale) and Pc promoters are indicated. Selected genes are shown by labeled arrows. Dotted lines joining arrowheads represent typical cassette array PCR products. (a) Tn*402*-like class 1 integron with the 5′-CS and a complete *tni* transposition region. (b) Typical class 1 integron structures with only part of the *tni* region and different extents of the 3′-CS. IS*1353* may also be present at the position indicated and IS*CR1* and one or more noncassette resistance genes may be inserted at the position indicated (after nucleotide 1313 of the 3′-CS). (c) Class 2 integrons are part of the large transposon Tn*7*. The asterisk represents the internal stop codon usually present in *intI2*. The *ybeA* gene is found within a cassette with an incomplete *attC* site. (d) The first class 3 integron characterized, in which *intI3* and cassettes are associated with a Tn*5053* family transposon.

the 3′-CS and only part of the Tn*402 tni* region. The 3′-CS is composed of several elements and different extents are present in different integrons. The start of the 3′-CS corresponds to the first 390 bp of *qacE*, the last cassette in Tn*402*, consistent with derivation from a Tn*402*-like transposon (Radstrom *et al.*, 1994). The truncated *qacEΔ1* gene overlaps with *sul1* (encoding sulphonamide resistance), which may also have once been part of a cassette (Stokes & Hall, 1989). Two ORFs of unknown function (orf5 and orf6) are found beyond *sul1* in some integrons and are also considered part of the 3′-CS. By convention, the end of the 3′-CS in each particular class 1 integron is defined as its boundary with another identifiable region. Common adjacent structures (Fig. 3b) include insertion sequences (IS) such as IS*1326* (with or without IS*1353*; 'In5-like') or IS*6100* flanked by inverted repeats of the end of the *tni* region ('In4-like'). In so-called 'complex' integrons, IS*CR1* and associated resistance genes are found between partial duplications of the 3′-CS (Toleman *et al.*, 2006).

The first few class 1 integrons to be identified were given integron (In) numbers intended to indicate the whole structure. For example, In2 carries the complete 5′-CS, the *aadA1a* gene cassette, 2025 bp of the 3′-CS followed by IS*1326*, IS*1353* and 2678 bp of *tni* including IRt (Liebert *et al.*, 1999). The conserved nature of the 5′-CS and that part of the 3′-CS immediately adjacent to the inserted cassette array enables amplification of the 'variable region' using a single primer pair (e.g. Lévesque & Roy, 1993; White *et al.*, 2000) (Fig. 3b). Many cassette arrays in class 1 integrons identified in this way have been given integron numbers, even though nothing is known about the structures beyond the start of the 3′-CS. This is counterproductive, as cassettes may move independently, or entire arrays move between integrons with different structures (Partridge *et al.*, 2002b). Even continuing to assign unique numbers to specific 'complete' integrons may not be helpful, as the epidemiology of class 1 integrons and gene cassettes appear to be somewhat independent.

Typical class 1 integrons lacking a complete *tni* region are defective transposon derivatives that are unable to move themselves, but if they retain both IRi and IRt they may still be transposed (e.g. Partridge *et al.*, 2002a). This transposition is presumably mediated by Tni proteins encoded by related intact transposons present in the same cell and is likely to happen only rarely. However, as Tn*402*-like transposons target the resolution (*res*) sites (Minakhina *et al.*, 1999) of plasmids (Kamali-Moghaddam & Sundstrom, 2000) and Tn*21*-like transposons (Liebert *et al.*, 1999), inserted integrons may then move as part of larger structures.

Integrons with *intI2*

The *intI2* gene was first described as part of the *c.* 14-kb transposon Tn*7* (Fig. 3c). Tn*7* includes the *tns* transposition region and is bounded by short segments, containing transposase-binding sites, called Tn7-L (*c.* 150 bp) and Tn7-R (*c.* 90 bp), which are necessary for transposition

(Peters & Craig, 2001). Tn*7* inserts at high frequency into a single specific site in bacterial chromosomes but is also able to transpose to many other sites at low frequency, with a marked preference for certain replicons on conjugative plasmids (Peters & Craig, 2001). Most examples of the *intI2* gene described to date contain an internal stop codon that renders IntI2 inactive, but natural suppression of the stop codon in IntI2 or the action of other IntI *in trans* (Hansson *et al.*, 2002) may allow occasional acquisition of new cassettes. In Tn*7* itself, the cassette array appears to end with a truncated cassette known as orfX or *ybeA* (Fig. 3c) and primers in *ybeA* and the conserved *intI2* region of Tn*7* (e.g. White *et al.*, 2001) have been used to amplify cassette arrays in class 2 integrons.

Integrons with *intI3*

In the first class 3 integron identified (Arakawa *et al.*, 1995) *intI3* is associated with a Tn*5053*-family transposon (Collis *et al.*, 2002) but in the opposite orientation compared with *intI1* in Tn*402*, with IRi found just beyond the end of the last cassette (Fig. 3d). *intI3* genes are generally not detected in surveys for *intI* genes in clinical isolates (e.g. Yu *et al.*, 2003).

Gene cassette nomenclature

Although the process by which gene cassettes are assembled remains unknown, it is possible that the same gene could become associated with different *attC* sites. However, this has generally not been observed and cassettes are traditionally given the same name as the gene they carry. Unfortunately, the naming of new genes/cassettes is not formally regulated and several authors have commented on the confusion caused by the same name being given to two different genes/cassettes (Vanhoof *et al.*, 1998; White *et al.*, 2000; Lee & Jeong, 2005). The same gene/cassette may also be given different names, the simplest examples being the use of Roman vs. Arabic numerals or an updated gene name that is not universally adopted (e.g. *dhfrVII* vs. *dfrA7* for a dihydrofolate reductase; see Table 1).

In other cases, alternative nomenclature systems are well established. One scheme for genes encoding aminoglycoside modifying enzymes is based on the type (acetylation, *aac*: adenylylation, *ant*; phosphorylation, *aph*) and site (3, 3′, 6′, etc.) of modification and the resistance profile (I, II, etc.), with a, b, etc. used to distinguish unique proteins (Shaw *et al.*, 1993). This system is becoming more difficult to apply, as new genes encoding potential aminoglycoside-modifying enzymes (especially those sequenced as part of large plasmids) are increasingly recognized and named on the basis of sequence similarity alone. An alternative scheme based on guidelines for naming plasmid genes (Novick *et al.*, 1976) is often used for cassette-borne aminoglycoside resistance genes and has been used here. In this system, *aacC* equates with *aac(3)*, *aacA* with *aac(6′)*, *aadA* with *ant(3″)*, *aadB* with *ant(2″)* and *aphA* with *aph(3′)*, with 1, 2, 3, etc. distinguishing different genes/gene cassettes.

Each cassette may also have several minor variants and deciding when a variant cassette should be given a separate name is problematic. In some cases (e.g. β-lactamases) a single amino acid change can dramatically change the resistance phenotype and is considered sufficient for assigning a unique protein number (http://www.lahey.org/Studies/), but cassettes encoding the same protein can have silent nucleotide differences. For other gene families, often where the effects of minor sequence variations on resistance phenotype have not been investigated and/or are of little clinical importance, variants may not be given separate names.

Some gene cassettes identified in MRI include ORFs potentially encoding proteins of as yet unknown function. Those identified some time ago generally have well-established names (e.g. orfA and orfC), but these names are also likely to be used for other completely unrelated ORFs. More recently identified examples are often not annotated or are given generic names, frequently 'orfX' or 'orf1' (see Table 2). As well as creating confusion, some such cassettes may have functions other than to encode proteins (Holmes *et al.*, 2003) and we believe that an alternative to the 'orf' designation is necessary. We propose the term *gcu*, for gene cassette of unknown function, for sequences found in MRI that have a credible *attC* site but for which a function cannot yet be assigned. We have retained the established letter designation for cassettes up to orfQ i.e. orfA becomes *gcuA*, orfC becomes *gcuC*, etc., with variants of orfD and orfE (originally mostly called 'orfE-like') designated *gcuD1*, *gcuE1*, *gcuE2* etc. Other *gcu* were numbered in order of the date of the first GenBank submission.

Automated annotation of cassettes and cassette arrays

Several compilations of gene cassettes have been published (e.g. Recchia & Hall, 1995b; Fluit & Schmitz, 1999, 2004; Rowe-Magnus & Mazel, 2002). However, wide use of PCR to amplify cassette arrays (particularly from class 1 integrons) has resulted in a huge proliferation of sequences in GenBank and no current summary of resistance gene cassettes is available. The task of analysing the available sequence data to produce such a compilation is daunting due to the large number of relevant sequences and is hampered by inconsistent nomenclature and poor annotation. Although methods for automated analysis of DNA sequences have been developed in recent years, the regions containing mobile resistance genes in Gram-negative bacteria present specific problems. In contrast to other systems, identifying and assigning a function to a newly sequenced resistance gene is

Table 1. Gene cassettes carrying known antibiotic resistance and related genes

Name in FDB	Other names	Named variants*	GenBank accession number	Start	End	*attC*†	No.‡
Aminoglycoside (6′) acetyltransferases§							
aacA1:gcuG¶	*aac(6′)-Ia*		AF047479.2	2143	3300	118	6
aacA2	*aac(6′)-Id* orfB		X12618.1	896	1421	72	1
aacA3	*aac(6′)-IIa*		AY123251.1	2926	3553	60	12
aacA4	*aac(6′)-Ib*		U59183.1	281	919	72	204
aacA5	*aac(6′)-IIb*		L06163.1	507	1159	97	1
aacA7	*aac(6′)-II*		U13880.2	299	889	112	36
aacA8			AY444814.1	3328	3966	72	2
aacA16	*aac(6′)-II,m,p*		Z54241.1	421	1132	110	2
aacA17	*aac(6′)-Ip,q*		AF047556.1	78	789	109	3
aacA27	*aac(6′)-IIc*		AF162771.1	36	773	70	5
aacA28	*aac(6′)-Iae*		AB104852.1	1905	2551	62	3
aacA29	*aacA29a,b*		AY139599.1	737	1248	112	4
aacA30	*aac(6′)-I30*		AY289608.1	1472	2197	112	1
aacA31	*aac(6′)-31*		AJ640197.1	2414	3052	72	3
aacA32	*aac(6′)-32*		EF614235.1	2222	2918	129	1
aacA33-‖	*aac(6′)-30*		AJ584652.2	1914	2334	–	1**
aacA34	*aac(6′)*		AY553333.1	1376	1841	58	3
aacA35	*aac*		AJ628983.2	1960	2659	132	1
aacA37	*aac(6′)*		DQ302723.1	69	530	60	2
aacA39	*aac(6′)-Iai*		EU886977.1	514	1207	109	1
aacA40††	*aac(6′)-IIa*		EU912537.1	2067	2766	132	1
Aminoglycoside (3) acetyltransferases‡‡							
aacC1	*aac(3)-Ia*		U90945.1	2568	1992	109	28
aacC2§§	*aac(3)-Ib*		L06157.1	609	1106	–	1**
aacC3	*aac(3)-Ic*		AJ511268.1	1276	1865	112	2
aacC4	*aacC1*		AF318077.1	2020	2595	108	2
aacC5	*aac(3)-Id*		AB114632.1	83	646	78	13
aacC6	*aac3-I*		AY884051.1	49	625	112	2
aacC7¶¶	*aacC-A7*		CP000282.1	2 333 601	2 334 159	75	1
aacC11	*aac(3)-I*		AJ877225.1	5772	5198	107	1
Aminoglycoside (3″) adenylyltransferases (streptomycin/spectinomycin resistance)							
aadA1a	*ant(3″)-1a*		X12870.1	1290	2145	60	259
aadA2			X68227.1	145	1000	60	150
aadA4			AF364344.1	756	1649	55	3
aadA5			AF137361.1	55	949	57	63
aadA6	*aadA11*		AF140629.1	52	914	60	15
aadA7			AF224733.1	23	884	60	14
aadA9			AJ420072.1	26 764	27 664	60	8
aadA10			FM207632.1	618	1480	60	1
aadA11	*aadA11b*		AM261282.1	23 611	24 466	60	1
aadA13			AY940492.1	72	926	60	8
aadA16	*ant(3′)-Ij*		EU678897.1	1684	2546	60	4
aadA24			AM711129.1	1243	2098	60	1
Aminoglycoside (2″) adenylyltransferases							
aadB	*ant(2″)-1a*		L06418.4	1287	1877	60	89
Aminoglycoside (3′) phosphotransferases							
aphA15			Y18050.2	4748	5659	108	15
Class A β-lactamases							
blaP1	PSE-1,CARB-2	*P2*‖‖, PSE-4,5 CARB-3,8	Z18955.1	102	1145	111	39
blaP3	CARB-4		U14749.1	623	1680	126	3
blaP7¶¶	CARB-7	*P9* (CARB-9)	AF409092.1	819	1879	128	2
bel1			DQ089809.1	959	1924	63	1
ges1	IBC-1	2,3,4,5,6,7,8,9	AF355189.1	2500	3519	110	10
veb1	CEF-1	2,3,4,5,6	AF010416.1	129	1198	133	13
Class B metallo-β-lactamases							
imp1	ESP	3,6,10	D50438.1	1179	2058	127	21
imp2		8,19,20,24	AJ243491.1	523	1353	78	6
imp4			AF445082.1	1344	2229	133	4
imp5†††			AF290912.1	< 1	> 741	132	1**
imp7			AF318077.1	1132	2019	135	2
imp9		25	AY033653.3	2771	3656	131	2
imp11		21	AB074436.1	1	885	131	2
imp12	IMP-10		AJ420864.1	140	1023	131	1
imp13			AJ550807.1	2227	3112	133	4
imp14			AY553332.1	98	983	133	3
imp15			AY553333.1	98	991	141	2

Table 1. Continued.

Name in FDB	Other names	Named variants*	GenBank accession number	Start	End	*attC*†	No.‡
imp16			AJ584652.2	1028	1913	133	1
imp18			EF184215.1	467	1354	135	1
imp22			DQ361087.2	340	1225	133	1
vim1		4,14	Y18050.2	3195	4108	81	41
vim2		3,6,8,9,10,11,15,16,17,18	AF191564.1	1286	2194	72	67
vim5			DQ023222.1	1286	2199	81	1
vim7			AJ536835.1	121	1036	87	2
vim13			EF577407.1	1151	2082	99	1
gim1			AJ620678.1	1022	1797	56	1
sim1			AY887066.1	478	1349	88	3
Class D β-lactamases							
oxa30†††	OXA-1	31,33	AF255921.1	1290	2293	90	35
oxa2		15,32,34,36,102	M95287.4	2435	3310	70	46
oxa5			X58272.1	63	977	106	2
oxa9			M55547.1	2269	3225	69	8
oxa10	PSE-2	11,14,16,17,74,142	U37105.2	1287	2206	111	7
oxa13		4,7,19,28,35,56	U59183.1	920	1791	63	10
oxa20		37	AF024602.1	1217	2169	117	5
oxa21		3	DQ522237.1	3953	4828	70	2
oxa46			AF317511.1	2350	3211	56	2
oxa53			AY289608.1	596	1471	70	1
oxa118			AF371964.1	98	958	55	3
oxa129			AM932669.1	682	1598	108	1
Chloramphenicol acetyltransferases							
catB2			AJ487034.1	1544	2282	72	9
catB3			U13880.2	890	1604	60	39
catB5			X82455.1	2	714	60	3
catB6			AJ223604.1	2986	3715	77	1
catB8			AY123251.1	780	1492	60	12
catB9¶¶			AF462019.1	1	772	128	2
catB10			AJ878850.1	1163	1877	60	1
catB11‡‡‡	cat-like		DQ831140.1	4600	5354	93	2
Chloramphenicol exporters							
cmlA1		4,5,6,7	U12338.3	6736	8284	70	48
cmlA2	*cmlB*		AF034958.3	2436	3983	70	1
cmlA8			EU182575.1	1480	3028	70	3
Dihyrofolate reductases (trimethoprim resistance)							
dfrA1	*dhfrIb dfr1 dhfrI*		X00926.1	216	792	95	162
dfrA5	*dhfrV dfrV*		X12868.1	1287	1854	87	13
dfrA6¶¶	*dfrVI*		Z86002.1	317	923	126	1
dfrA7	*dhfrVII dfrVII dfrA17*		X58425.1	573	1189	134	22
dfrA12	*dhfrXII dfr12*		Z21672.1	302	885	90	52
dfrA14	*dhrfIb*		Z50805.1	53	619	86	14
dfrA15	*dhfrXVb*		Z83311.1	337	922	104	19
dfrA16	*dhfrXVI dfr16*		AF174129.3	1333	1920	107	11
dfrA17	*dhfrXVII dfr17*		AF169041.1	141	756	133	54
dfrA21§§§	*dfrxiii*	13	AJ870926.1	5508	6091	90	4
dfrA22	*dfr22 dfr23*		AJ968952.1	237	820	90	2
dfrA25			DQ267940.1	20	590	90	3
dfrA27	*dfr*		EU678897.1	1021	1582	82	4
dfrA28			FM877476.1	98	659	82	1
dfrA29¶¶¶	*dfrVII dfrA7*		AM237806.1	594	1209	133	2
dfrA30‖‖‖	*dhrfV*		AM997279.1	624	57	87	1
dfrA31¶¶,****	*dfr6*		AB200915.1	2324	1718	126	1
dfrB1	*dhfrIIa dfr2a dfrII*		AY139601.1	98	508	57	6
dfrB2	*dhfrIIb dhfr*		J01773.1	707	1090	57	6
dfrB3	*dhfrIIc dfr2c*		U67194.4	33 061	33 468	57	4
dfrB4	*dhfr2 dfr2d*		AJ429132.1	2	409	57	7
dfrB5	*dhfrB5 dfrIIe dhfr2e*		AY943084.1	2786	3196	57	8
dfrB6			DQ274503.1	325	734	57	1
dfrB7††††	Not annotated		DQ993182.1	64	540	72	1
Streptothricin acetyltransferases							
sat2‡‡‡‡			X15995.1	247	830	60	56
Quaternary ammonium compound efflux							
qacE			U67194.4	33 789	34 375	141	2
qacF			AF034958.3	1926	2435	60	8
qacG	*qacE2,G2*		AF327731.1	10	546	99	8

Table 1. Continued.

Name in FDB	Other names	Named variants*	GenBank accession number	Start	End	*attC*†	No.‡
qacI§§§§	*qacH*		AF205943.1	1260	1770	60	13
qacK¶¶¶¶			EF522838.1	3564	4127	117	1
Small multidrug resistance proteins							
smr1			AF406792.1	141	546	62	1
smr2	*smr* orfO		AY260546.3	5455	5859	60	7
ADP-ribosyl transferases (rifampicin resistance)							
arr2		3	AF078527.1	4352	4954	114	26
arr4-R‖			EF660562.1	1631	2146	-	1**
arr5			EF660563.1	352	906	60	1
Erythromycin esterases							
ereA1			AY183453.1	2657	4025	57	4
ereA2			AF099140.1	62	1437	57	6
Lincomycin nucleotidyltransferases							
linF			AJ561197.2	1247	2181	58	1
linG			DQ836009.1	1234	2170	58	3**
Fosfomycin resistance							
fosB-A‖,¶¶			X54227.1	684	1140	–	3
fosE	orfI, orfi		AY029772.1	2193	2654	60	2
fosF	orf1		AY294333.1	495	1005	93	1
fosG	ORFV fosC		AY907717.1	2973	3469	78	2
fosH	orf2, orf2a		DQ342344.1	1406	1872	58	4
Quinolone resistance							
qnrVC1			EU436855.1	1186	2197	126	1
qnrVC2¶¶			AB200915.1	3216	2325	129	1

*Amino acids changes in variants of β-lactamase cassettes are listed in Tables S1–S8 and for *ges* in Table 9.

†The length of the *attC* site (this may differ slightly in variants).

‡The number of GenBank entries with a complete version of the gene cassette is given, including examples in all contexts (MRI, chromosomal integrons and secondary sites).

§Early examples of the *aac(6′)* group were designated *aac(6′)-I* or *aac(6′)-II* on the basis of resistance phenotype but this does not always reflect genetic relatedness and these genes fall into several clusters. Here, all cassettes of this type have been given an *aacA* number, corresponding to the letter in the *aac(6′)* name where possible. Low numbers that appear not to have been used for other genes have been used for cassettes first identified some time ago and numbers > 26 have been used for more recently identified cassettes.

¶*aacA1*:*gcuG* contains two ORFs but only one *attC* site. The whole sequence was used as a feature and all partial examples identified here were due to the sequence in GenBank starting or ending within the cassette.

‖The only available versions of these cassettes have precise truncations in the *attC* site. In the case of *aacA33*, it is not possible to tell whether the 7 bp at the end correspond to the left or right spacer.

**The number of GenBank entries with incomplete versions of the cassettes is given, as the complete versions have not yet been identified.

††The cassette gene was annotated as *aac(6′)-IIa* in EU912537 but the protein is only 78% identical to AAC(6′)-IIa.

‡‡*aacCA* (Levings *et al*., 2005) or *aacC-A* (Elbourne & Hall, 2006) have also been used for this gene family.

§§Part of the *attC* site is missing from the cassette sequence.

¶¶These cassettes have only been identified outside an MRI context to date. It is possible that other antibiotic resistance cassettes that are only found outside an MRI context were not identified here.

‖‖*blaP2* is listed as a separate cassette in previous compilations, but is > 98% identical to *blaP1*.

***Only the sequence of the *imp5* gene is available in GenBank accession number AF290912. The full cassette sequence is from Da Silva *et al*. (2002).

†††The original sequence of the *oxa1* cassette was found to contain an error (Boyd & Mulvey, 2006) and the correct sequence matches *oxa30*.

‡‡‡The cassette was annotated as encoding a 'Cat-like protein' that is 87% identical to CatB2.

§§§*dfrA13* was identified first, but has probable errors compared with the related *dfrA12* cassette; hence, the *dfrA21* sequence was used here.

¶¶¶This cassette was annotated as *dhfrVII* in AM237806 and called *dfrA7* in O'Mahony *et al*. (2006) but is only 80% identical to the *dfrA7* gene cassette.

‖‖‖This cassette was annotated as *dhfrV* in AM997279 but is only 93% identical to the *dfrA5* cassette.

****The gene was annotated as *dfr6* in AB200915, but the cassette is only 90% identical to the *dfrA6* cassette.

††††Neither a cassette nor a gene was annotated in DQ993182. The protein is 85% identical to DfrB2.

‡‡‡‡The *estX* cassette (see Table 2) is sometimes incorrectly identified as *sat* (Partridge & Hall, 2005).

§§§§This cassette is annotated as *qacH* in GenBank AF205943, but called *qacI* in the accompanying publication (Naas *et al*., 2001). *qacI* has been used here, as there is a distinct *qacH* gene in *Staphylococcus*.

¶¶¶¶Neither a cassette nor a gene was annotated in EF522838. The protein is 81% identical to QacE and was designated *qacK*, as *qacJ* had already been assigned to a distinct *qac* gene in *Staphylococcus aureus*.

Table 2. Gene cassette in MRI not encoding known resistance genes

Name in FDB	Other names/annotations	GenBank accession number	Start	End	*attC*[*]	No.[†]
estX	*sat*[‡]	AB121039.1	65	1016	71	30
psp[§]		AB121039.1	1017	1689	60	14
lsp[¶]		EU780012.1	2721	3331	111	1
gcuA	orfA	U12441.2	3371	3871	69	2
gcuC	orfC orfX	AF455254.1	650	1161	60	35
gcuD	orfD	M95287.4	5043	5362	60	40
gcuD1	orf4; similar to orfD	DQ278189.1	85	403	60	1
gcuE	orfE	U12338.3	5618	5879	60	2
gcuE1[‖]	orf3				60	0
gcuE2	ORF1 orfE like	AJ487033.2	689	950	60	5
gcuE3	orfE like	AY139595.1	1248	1509	60	1
gcuE4	orfE like	AY139597.1	98	360	60	1
gcuE5	orfE like	AJ564903.1	16 579	16 321	57	3
gcuE6	orf1	AY758206.1	2452	2734	60	1
gcuE7	Similar to orfE like	DQ522236.1	1422	1740	59	2
gcuE8	orfE like	EU434616.1	320	58	60	1
gcuF	orfF	Z21672.1	886	1205	60	53
gcuF1	Similar to orfD[**]	FJ207466.1	1317	1635	60	1
gcuH[††]	orfH	AF047479.2	3301	3752	86	1
gcuI[††]	orfI	AF047479.2	3753	4350	77	1
gcuJ[††]	orfJ	AF047479.2	4351	4726	74	1
gcuN	orfN orfI	AJ223604.1	3716	4404	75	1
gcuO	ORFO ORFX ORF	AJ251519.1	2	467	76	5
gcuP	orfX orfXA orf9	U90945.1	1991	1454	70	21
gcuQ	orfX′ orfY orfXB orf10	U90945.1	1453	1055	69	21
gcu1	Not annotated	AF318077.1	98	492	117	1
gcu2	orf2a, orf2b	AY139592.1	1267	1606	88	2
gcu3	orf1	AY139600.1	1103	1505	72	1
gcu4	Not annotated	AJ536835.1	1037	1307	102	2
gcu5a[‡‡]	Unknown	AY220520.1	1163	1588	91	2
gcu5b[‡‡]	Not annotated	DQ520941.1	3295	3720	91	1
gcu6	orfA	AB113580.1	3465	4118	78	2
gcu7	orfPa85	AJ634050.2	3527	3899	60	1
gcu8a[‡‡]	orf416, ORF1	AJ704863.3	22 742	22 299	60	4
gcu8b[‡‡]	Not annotated	AJ487033.2	951	1393	60	1
gcu9[§§]	ORF1, hypothetical protein	AJ784256.1	1563	2203	76	2
gcu10	orfPa105	AJ786649.2	3842	4491	73	1
gcu11	Cassette without gene or orf	AB195796.1	2029	2599	71	4
gcu12	ORFX	AJ871915.1	1878	2407	60	1
gcu13	ORFIV, ORFVI, orfvi	AY907717.1	1113	1424	56	2
gcu14	ORF126	DQ236170.1	286	550	85	1
gcu15	Orf1	DQ278188.1	762	1166	69	1
gcu16	orf2, ORF IN682	DQ278190.1	32	468	104	2
gcu17	orfX	DQ361087.2	1226	1736	61	1
gcu18	Not annotated	AM237806.1	65	593	132	2
gcu19	GCN-5 acetyltransferase	AM237806.1	1210	1726	72	2
gcu20	orf1	DQ520941.1	98	589	90	1
gcu21	Not annotated	DQ522237.1	1940	2403	85	1
gcu22	ORF1 and ORF2	DQ533990.1	1444	2908	72	1
gcu23	ORF3 and ORF4	DQ533990.1	2909	4323	72	1
gcu24	ORF5	DQ533990.1	4324	5571	72	1
gcu25	ORF6	DQ533990.1	5572	6086	75	1
gcu26	ORF1	AM711129.1	645	1242	72	1
gcu27[¶¶]	Not annotated	EF614235.1	3828	5621	60	1
gcu28	Cassette, unknown function	EU165039.1	1512	1989	111	1
gcu29	Hypothetical protein	EU284133.1	204	651	60	1

Table 2. Continued.

Name in FDB	Other names/annotations	GenBank accession number	Start	End	*attC**	No.†
gcu30‖‖	orf102	DQ914960.2	208	645	85	1
gcu31	Not annotated	EU434611.1	936	574	79	1
gcu32	ORF1	EU487199.1	858	1374	132	1
*gcu33****	JK0007	EU591509.1	10 532	11 161	60	1
gcu34	ORF4	FM179467.1	2794	3193	76	1
gcu35	Unannotated	EU588392.2	2742	3057	59	1
gcu36	orfV	EU886977.1	1208	1486	60	1

*The length of the *attC* site (this may differ slightly in variants).

†Number of GenBank entries containing the complete cassette.

‡The *estX* cassette is often mistakenly identified as *sat*, but the gene encodes a putative esterase (Partridge & Hall, 2005).

§The *psp* gene encodes a putative phosophoserine phosphatase.

¶The *lsp* gene encodes a putative lipoprotein signal peptidase.

‖The *gcuE1* sequence is not available in GenBank and was obtained from Yano *et al.* (2001).

**This cassette is 75% identical to the *gcuF* cassette and 74% identical to the *gcuD* cassette.

††In AF047479 ORFs annotated as orfK, orfL and orfM follow the *gcuJ* cassette. The region after orfKL contains a potential 1L core site and could form a folded structure but the typical 2L/2R pairing is not evident. The region after orfM also contains a potential core site but an appropriately folded structure was not predicted. This region was included in the FDB as noncassette insertion (designated KLM).

‡‡*gcu5a* and *gcu5b* are 96% identical, as are *gcu8a and gcu8b*.

§§The ORF in *gcu9* may encode a quinolinate synthetase.

¶¶The only available example of the *gcu27* cassette is interrupted by ISUnCu*1*, positions 4199–5580 in EF614235.

‖‖The only example of *gcu30* is inserted in at a secondary recombination site in the 5′-CS.

***The ORF in *gcu33* encodes a predicted NADPH-dependent FMN reductase.

often relatively simple, as these genes often encode proteins belonging to easily recognizable families. However, identification of the boundaries of potentially mobile regions (e.g. gene cassettes) and consistent annotation are both extremely important.

We have developed a novel bioinformatics tool to enable a detailed analysis of gene cassettes and cassette arrays in MRI, to be described in more detail elsewhere (G. Tsafnat *et al.*, unpublished data). A database of defined sequence 'features' and automated BLASTN searches were used to identify and reannotate sequences containing gene cassettes. We then used similarities between DNA and natural languages (Baquero, 2004) to develop a context-sensitive grammar to define higher order genetic structures (cassette arrays) from annotations of 'features' (gene cassettes). While computational grammars have previously been used to decode genetic patterns at the letter-to-word (base pair to feature) level (e.g. Leung *et al.*, 2001; Searls, 2002), our grammar operates at the word-to-sentence level.

The feature database

We compiled a feature database (FDB) of gene cassettes ('features') listed in previous reviews and those reported more recently. Cassettes were identified by nucleotide/protein similarity searches, by searches of GenBank and PubMed with the word 'integron' and iteratively during the automated annotation process. Closely related cassettes (> 98% identical) were grouped, regardless of phenotype conferred, and an exemplar GenBank accession number chosen for each group. This was generally the first report, unless errors were apparent, or the most common sequence among minor variants. The span of the cassette (Tables 1 and 2) and the sequence were recorded in the FDB. A minimum percentage of base pair identity (usually 97%) required for part of a sequence to be considered a match was also recorded. Short sequences flanking cassette arrays in different MRI, i.e. the ends of the 5′-CS, 3′-CS and Tn*402 tni* of class 1 integrons, the *intI2* region and the *ybeA* cassette of class 2 integrons and the *int3* and IRi regions from the first class 3 integron, were also included as features.

A single name was selected to represent each cassette sequence in the FDB, but alternative names are also listed in Tables 1 and 2. Where possible the name given in the original GenBank entry or publication was used. For cassettes that were not suitably annotated in the original GenBank entry BLASTN or BLASTX searches were used to assign an appropriate gene family name and the next apparently available letter and/or number was used. For simplicity, cassettes for which the full name of the gene is of the form $bla_{\text{XYZ-1}}$ are represented as *xyz1* throughout.

Annotation of cassette arrays in MRI

All features in the FDB were used in a BLASTN (Altschul *et al.*, 1997) search against the complete nucleotide GenBank

database (http://www.ncbi.nlm.nih.gov) (Benson *et al.*, 2009). GenBank entries containing any segment that met the minimum identity match criteria for any feature in the FDB were collected. The species from which the sequence was obtained was acquired from the organism field and entries with the words 'vector', 'synthetic construct' or 'artificial sequence' were excluded, as were a few (e.g. DQ915900–DQ915939) containing long runs of Ns representing unsequenced regions. Sequences in the RefSeq collection (accession numbers of the form 'NC_', etc.) were also excluded, as each is derived from and has a sequence identical to an entry with a conventional accession number.

Sequences were annotated with the most similar features from the FDB (as determined by the BLAST bit score) without reference to annotations in GenBank. Many sequences carry 'partial' features, created either because the end of the submitted sequence lies within the feature or because of an insertion, leading to gaps in annotations. Compiling a database from all sequences in the FDB and searching with the sequences of these gaps allowed partial features to be annotated; these were designated by # after the feature name. Allowing the system to register matches of < 25 nt introduced an unacceptable number of false annotations, but many sequences of 'cassette array PCR' amplicons include < 25 nt flanking sequence at one or both ends. The automated annotation process also missed a few short 5′-CS and/or 3′-CS that were > 25 nt but contained a significant number of differences (possibly errors) from the standard sequences. In these cases, the sequence and, if possible, the corresponding paper were checked and annotations of short flanking regions were added manually as appropriate.

A context-sensitive grammar (Grune & Jacobs, 2007), consisting of 21 rules to identify cassette arrays flanked by end markers, was then applied to parse cassette arrays (in either orientation in sequences in GenBank) from annotated features. The parser also examined the context of annotation gaps, identifying those found within an array as potential cassettes that were missed when the FDB was first compiled. These were checked manually and, if appropriate, added to the FDB as gene cassettes. Other elements identified within cassette arrays were added to the FDB as 'noncassette insertions' and were registered by the grammar as features that do not 'break' a cassette array. These include group II introns, IS, short regions possibly corresponding to the beginning of a truncated cassette [designated 'potential cassette starts' (PCS)] and rare longer inserts of unknown origin.

Data were presented as a searchable list of all collected GenBank entries with the names and the spans of all annotated features indicated and any complete cassette arrays identified. The number of complete and partial versions of each cassette and the composition and number of each cassette array found in the different MRI were automatically compiled. In the following sections, we give a summary of gene cassettes found in sequences of MRI lodged in GenBank. We also list common modifications of cassettes and describe selected illustrative cassettes in more detail. The aim is to provide a reference tool, including extensive tables that will be periodically updated online (http://www2.chi.unsw.edu.au/genecassettes).

Gene cassettes and cassette arrays

Gene cassettes

All resistance gene cassettes identified here and used as features in the FDB are listed in Table 1, grouped by resistance conferred and gene type, with *gcu* in Table 2. If $< 98\%$ identity is used as the cut-off for defining a new cassette, 132 different cassettes carrying known antibiotic resistance genes (or homologues assumed to confer similar resistance phenotypes) and 62 different *gcu* were found in MRI. If the same criteria are used, this compares with 40 cassettes (35 resistance gene cassettes+5 *gcu*) compiled in 1995 (Recchia & Hall, 1995b), 53 (47+6) in 1999 (Fluit & Schmitz, 1999), 69 (63+6) in 2002 (Rowe-Magnus & Mazel, 2002) with 7 (4+3) added in 2004 (Fluit & Schmitz, 2004) to give 76 (67+9), indicating that novel cassettes continue to be acquired by MRI.

While most cassettes identified since the last compilation belong to one of the known families, cassettes conferring resistance to fosfomycin (Yatsuyanagi *et al.*, 2005) and lincomycin (Heir *et al.*, 2004) and, most recently, presumed to confer quinolone resistance (Fonseca *et al.*, 2008) have now been found. We identified several *gcu* that were not annotated in the original GenBank entries (Table 2) as well as two novel cassettes carrying putative antibiotic resistance genes, which we designated *dfrB7* (DQ993182) and *qacK* (EF522838). In several cases a resistance gene or cassette was annotated in the relevant GenBank entry, but the sequence was at most 93% identical to the exemplar with the same name (Table 1) and these were also designated as new cassettes.

attC sites

The sequences of all the *attC* sites of different cassettes identified here were compiled and their lengths are indicated in Tables 1 and 2, but presentation of a detailed analysis is beyond the scope of this review. Compilation of the *attC* sites from 39 cassette sequences available in 1997 suggested the consensus **GTT**AGSC/GYTCT**AAC** (top strand, completely conserved residues in bold) for 1R/1L (Stokes *et al.*, 1997) but GTTRRRY/RYYYAAC is also commonly used. Analysis of 1R sequences identified here indicated that GTTAGGC, GTTAGCC and GTTAGAC dominated. While, as expected, the fourth position was most commonly A, or

to a lesser extent G, C has now been seen at this position in one example (*gcu13*, GTTCTGT). The final nucleotide of 1R was A or G, rather than T or C, in a number of *attC* sites but few had mismatches between 1R and 1L.

Secondary structures of bottom strands were generated (http://mfold.bioinfo.rpi.edu/applications/hybrid/quikfold.php; Markham & Zuker, 2005, 2008) to help identify 2L and 2R and extrahelical bases. 2L and 2R sites were quite variable and difficult to identify conclusively for some *attC* sites. The extrahelical base in 2L was mostly G (*c.* 67%) or C (*c.* 25%) while A, and particularly T, were rarer. Where a second extrahelical base could be identified it was commonly T.

The shortest *attC* sites (55 nt) identified here were those of the *aadA4* and *oxa118* cassettes (two and three complete examples, respectively), while *attC* sites of the 'classical' 60 nt type (group 1 in Recchia & Hall, 1997) appeared most common. These 'classical' *attC* sites generally have G39 and T32 as extrahelical bases and several have been tested for activity, with the *aadB attC* reported as being highly active (Hall *et al.*, 1991) and *aadA1a* and *aadA7* were used in mechanistic experiments (Francia *et al.*, 1999; Johansson *et al.*, 2004; Bouvier *et al.*, 2005). The *gcuD* and *gcuF attC* sites are 60 nt and were classed as part of group 1, but have G31 rather than T32 as the second extrahelical base and the *gcuD attC* site appears to be active (Hall *et al.*, 1991). The *dfrB1-6 attC* sites are only 57 nt and have an extrahelical G on the top strand opposite T32, as noted by Hall *et al.* (1991).

The bottom strands of some longer *attC* sites are predicted to have more complicated secondary structures and the second extrahelical base may be more difficult to identify. The longest *attC* sites (up to 141 bp, group 3 in Recchia & Hall, 1997) include those of the *dfrA7*, *qacE* and most *imp* cassettes. These *attC* sites appear to have a loop immediately after the 2L/2R complementary region (Recchia & Hall, 1997). The *dfrA7 attC* was found to be active in cointegration assays but its activity seemed lower than most others tested (Collis *et al.*, 2001).

The typical *attC* in the *V. cholerae* chromosomal integron (VCR) has an additional extrahelical T at position 16″ (Demarre *et al.*, 2007). Although cassettes with this type of *attC* can be integrated/excised by IntI1 (Rowe-Magnus *et al.*, 2002) only two examples, *blaP3* and *qnrVC1*, have been seen in class 1 integrons, with *blaP7*, *catB9*, *dfrA6*, *dfrA31* and *qnrVC2* only identified in *V. cholerae* chromosomal integrons to date. Some shorter *attC* sites, including those of the related *oxa2*, *oxa21* and *oxa53* cassettes (70 nt) also appear to have a third extrahelical base.

The sequences of a few other *attC* sites were also more difficult to fit to the expected pattern. For example, the *veb1 attC* is not closely related to that of any other cassette identified here and in this case altering spacer lengths appears to be necessary to give the best match between potential 2L and 2R sites.

Cassette arrays flanked by the 5′-CS and 3′-CS

Over 300 different complete cassette arrays flanked by the 5′-CS and 3′-CS were identified in GenBank, most of which were found only once (Fig. 4a). Most arrays had two or three gene cassettes (Fig. 4b), but this may partly reflect a bias against PCR amplification of longer arrays. The most frequently lodged arrays (Table 3) generally correspond to those commonly found in surveys (e.g. Yu *et al.*, 2003; Machado *et al.*, 2007). Some of the common arrays would yield cassette PCR products of similar size (Table 3), but these may be distinguishable by restriction digests.

Most cassette array sequences in class 1 integrons with the 5′-CS and 3′-CS deposited in GenBank were from *Escherichia coli* (21%), *Pseudomonas aeruginosa* (19%), *Salmonella* spp. (14%), *Acinetobacter baumannii*, *Klebsiella pneumoniae* or *V. cholerae* (each *c.* 6%). Class 1 integrons carrying a few different cassette arrays have been reported in Gram-positive bacteria. These include *Corynebacterium* (Nesvera *et al.*, 1998), *Enterococcus* (Clark *et al.*, 1999) and, most recently, *Staphylococcus* spp. (e.g. Shi *et al.*, 2006).

Cassette arrays flanked by the 5′-CS and a complete *tni* region

A recent publication lists several cassette arrays flanked by the 5′-CS and the *tni* region (Post *et al.*, 2007). Three additional examples of the |*aacA7*|*vim2*|*dfrB5*|*aacC5*| array (AM749810-11, FM165436) and several additional arrays were identified here: |*imp4*|*qacG*|*aacA4*|*aphA15*|

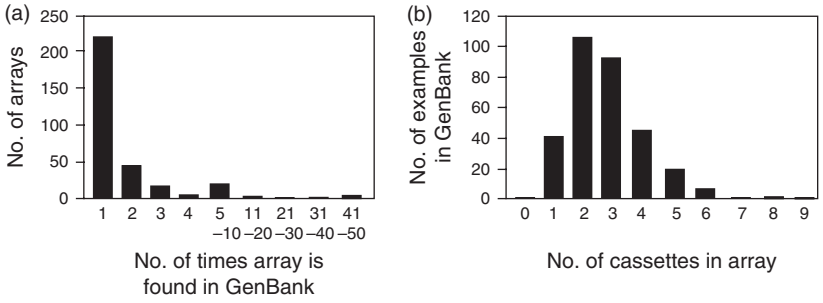

Fig. 4. (a) The frequencies of different cassette arrays flanked by the 5′-CS and 3′-CS in GenBank. (b) The number of cassettes in arrays flanked by the 5′-CS and 3′-CS.

(AF288045.2), |*dfrB1*|*aacA4*|*vim2*| (AM993098), |*aadB*|*qacI*-ISKpn*4*a-*qacI*| (EF408254), |*gcu33*|*qacG*| (EU591509) and |*vim2*|*gcu9*| (FJ237530). The complete Tn*402 tni* region (4733 bp) is found in plasmids R751 (Tn*402* itself; U67194) and pTB11 (AJ744860). In AY033653 the Tn*402 tni* region is truncated by another transposon and an almost complete version in AM993098 includes a short section only 85% identical to Tn*402*. EU591509 (Labbate *et al*., 2008) and AF288045 each include a *tni* region that is a hybrid of Tn*402* and a related transposon, with the crossover in the vicinity of the *res* site at position 792. All other entries end after < 700 bp of Tn*402 tni* sequence.

The apparent rarity of class 1 integrons with *tni* in place of the 3′-CS may reflect surveillance bias, but structures with the 3′-CS are likely to be at a selective advantage due to the presence of *sul1* (resistance to sulphonamides). The potentially less-biased set of class 1 integrons from complete resistance plasmid sequences in GenBank includes only two with the Tn*402 tni* region (in R751 and pTB11), compared with at least 30 with part of the 3′-CS.

Cassette arrays with the 5′-CS but no 3′-CS or *tni* region

Surveys for cassette arrays usually identify some *intI1*-positive isolates from which no amplicon is obtained with standard primers in the 5′-CS and 3′-CS (e.g. Yu *et al*., 2003). One possible explanation is failure to amplify long arrays containing many cassettes or large insertion(s) under commonly used conditions. In other cases *intI1* may be associated with the complete Tn*402 tni* region or may be found outside the Tn*402* context (Stokes *et al*., 2006; Gillings *et al*., 2008). A 'hybrid' integron in which *intI2* and the 3′-CS flank the cassette array has been reported (AJ289189) (Ploy *et al*., 2000) and the reciprocal structure with 5′-CS and the *tns* region may also occur, although no examples were identified by our methods. Recently, a region containing a transposase-like gene, commonly annotated as IS*440*, and the *sul3* sulphonamide resistance gene has been found beyond arrays that include *qacI* as the last cassette (Table 4). All sequences of this structure currently in GenBank are from *E. coli* or *Salmonella* and the cassette arrays are related, suggesting an ancestral structure with subsequent recombination within cassettes. Long-range PCR pairing a primer in the 5′-CS with one in the IS*440* region or *sul3* might enable amplification of these arrays.

Examination of other sequences in GenBank that contain the 5′-CS and a cassette array but lack the 3′-CS revealed several other possible explanations involving various IS. IS*6100* is found beyond the 3′-CS in In4-like integrons (Partridge *et al*., 2001) and IS*6100*-mediated deletions into the cassette array may explain some structures. In other cases, the array is truncated by IS*26* or IS*1*, both common components of multi-resistance regions and resistance plasmids. We have detected cassette arrays truncated by IS*6100* using a primer in the 5′-CS paired with a reverse primer in this IS (unpublished data). Similar pairings with primers in IS*1* or IS*26* (both orientations) may also yield products for some isolates with *intI1* from which a standard cassette array amplicon is not obtained.

Table 3. Common cassette arrays flanked by the 5′-CS and 3′-CS

Cassette array	Size (bp)*
|*aadB*|	591
|*dfrA7*|	617
|*aadA1a*|	856
|*aadA2*|	856
|*blaP1*|	1044
|*dfrA1*|*gcuC*|	1089
|*dfrA1*|*aadA1a*|	1434
|*dfrA17*|*aadA5*|	1511
|*dfrA12*|*gcuF*|*aadA2*|	1760

*Size does not include any 5′-CS or 3′-CS sequence.

Cassette arrays flanked by *intI2* and *ybeA*

The most commonly identified array flanked by *intI2* and *ybeA* was |*dfrA1*|*sat2*|*aadA1a*| ($n = 31$), as seen in Tn*7* itself, followed by |*estX*|*sat2*|*aadA1a*| ($n = 7$), |*sat2*|*aadA1a*| ($n = 4$), |*dfrA1*|*sat2*| ($n = 3$) and |*sat2*|*ereA1*|*aadA1a*| ($n = 1$). An additional array in *A. baumannii* with an unusual structure, |*sat2*|*aadB*|*catB2*#^|*dfrA1*|*sat2*|*aadA1a*| (DQ176450), where ^ represents the last 258 bp of the *intI2* region, is proposed to have arisen by integrase-mediated intermolecular recombination (Ramirez *et al*., 2005). Where the context has been examined, these arrays generally appear to be associated with Tn*7 tns* genes, but some, for example a |*dfrA1*|*sat2*|*aadB*| array with IS*911* inserted in *sat2* (EU732664)

Table 4. Cassette arrays flanked by the 5′-CS and the IS*440-sul3* region

Cassette array*	No.	GenBank
|*estX*|*psp*|*aadA2*|**IS*440*Δ**	1	CP000971
|*dfrA12*|*gcuF*|*aadA2/1*|*qacI*|**IS*440-sul3-IS26***	1	EF051038
|*dfrA12*|*gcuF*|*aadA2*|*cmlA1*|*aadA1a*|*qacI*|**IS*440-sul3-IS26***	1	EF051037
|*dfrA12*|*gcuF*|*aadA2*|*aadA2*|*cmlA1*|*aadA1a*|*qacI*|**IS*440-sul3-IS26***	1	EF113389
|*estX*|*psp*|*aadA2*|*cmlA1*|*aadA1a*|*qacI*|**IS*440-sul3-IS26***	7†	e.g. AY509004

**aadA2/1*, hybrid *aadA2/aadA1a* cassette.

†One example (in FJ160769) has IS*26* inserted in *aadA1a*, while the *cmlA1* cassette in EU219534 has a 331-bp deletion.

(Gassama Sow *et al.*, 2008), appear to be in a different context.

Arrays associated with *intI2* were reported from several different species, most commonly *E. coli* (16/48) and *Shigella* (11/48), but these integrons are relatively unimportant clinically and are likely to remain so unless they acquire more varied cassettes by capture or recombination. However, the recent finding of two examples of an *intI2* gene without the usual internal stop suggests that cassette arrays in class 2 integrons should continue to be monitored. One intact *intI2* is associated with a *dfrA14* cassette and a possible lipoprotein signal peptidase (*lps*) with no additional context information available (Márquez *et al.*, 2008), the other is associated with several *gcu* and the Tn*7 tns* genes (Barlow & Gobius, 2006).

Cassette arrays associated with *intI3*

Only two different *intI3*-associated cassette arrays that include known resistance genes were detected in GenBank. The |*imp1*|*aacA4*| cassette array is found in D50438, AB070224 and AF416297, all of which were obtained from the same plasmid from *S. marcescens* (Arakawa *et al.*, 1995; Collis *et al.*, 2002). PCR with primers for *intI3*, *imp1* and *aacA4* detected similar integron structures in isolates of other species from Japan (Senda *et al.*, 1996). The second array, |*ges1*|*oxa10*:*aacA4*|, was identified in Portugal (AY219651) (Correia *et al.*, 2003). The sequence beyond *intI3* is not available in this case but a *rep* gene related to that of plasmid RSF1010 was identified beyond the cassette array, rather than the region containing IRi of a Tn*402*-like transposon seen in the isolate from Japan.

Cassettes in secondary sites

A few cassettes found in MRI were also found in secondary integration sites (Recchia & Hall, 1995a). These included *aadB* in plasmid backbones (U14415, AF003958) and *dfrA14* in the *strA* gene (e.g. AJ313522). The only example of the *gcu30* cassette is inserted within the *intI1* gene (DQ914960).

Detailed analysis of selected cassettes and arrays

A number of cassettes with modifications were identified here. Many of these modifications were overlooked in the original GenBank entry and/or publication, as sequence analysis and annotations are often limited to cassette-borne genes, rather than complete cassettes. These modifications may have implications for cassette movement or expression of cassette genes and/or be useful as epidemiological markers, as well as complicating nomenclature. In the sections below we provide instructions for identifying cassette boundaries and *attC* sites, and give examples of common types of cassette modifications.

Finding the boundaries of a cassette and correctly identifying the *attC* site

Identifying cassette boundaries and *attC* sites is useful for analysing array sequences (allowing separate searches with each component cassette), for annotation and for identifying modifications. For an array in a class 1 integron (see Fig. 5a) searching with aaaacaaagTT usually allows identification of the boundary between the end of the 5′-CS (the g) and the start of the first cassette (the first T, hereafter position 1). The gTT and the following four bases give the sequence of the 7 nt 1R core site. In most cases (e.g. the first cassette in Fig. 5a) searching for the reverse complement of 1R identifies 1L, corresponding to the first 7 nt of the *attC* site. A gap of 5 nt separates the final C of 1L from the start of the 8 nt 2L core site. In the case shown, taking the sequence of 2L, removing the fourth base (usually C) and searching with the reverse complement identifies the 7 nt 2R core site. A gap of 5 or 6 nt separates the end of 2R from the G that defines the end of the first cassette and of its *attC* site. The t following this G defines the start of the next cassette and the steps above can be repeated to identify the end and *attC* site of this and following cassettes. The start of the 3′-CS is defined by TTAGAT, but all *qac* cassettes also begin with this sequence.

In some cassettes (e.g. the second cassette in Fig. 5a) 1R and 1L are not completely complementary beyond GTT/AAC. Notable examples are the *aadA1* (GTTA**A**AC/GT**C**TAAC), *gcuE*-like (GTTAG**T**C/G**T**CTAAC) and *gcuF* (GTTAG**C**A/T**T**CTAAC) cassettes. In such cases marking ORFs and searching for YAAC close to stop codons may allow identification of 1L. 1L sites ending in TAAC are most common and the TAA triplet often corresponds to the stop codon of the cassette gene. However, the ORF may also end before 1L (e.g. *aadA1a*), extend further into the *attC* site (e.g. *aadA10*) or even continue right through the *attC* site, ending within the 1R site of the next cassette (e.g. *gcuD*, *gcuE*, *gcuF* and related cassettes).

Cassettes (e.g. the second example in Fig. 5a) may also have mismatches between 2L and 2R. In these cases searching for Gttr sequences *c.* 40–130 nt beyond the end of 2L but before the start of the next ORF (up to *c.* 300 bases from position 1 in known cassettes) identifies possible boundaries with the next cassette. Counting back six or seven positions from the g of these gTTR motifs will generally identify the eighth nucleotide of potential 2R sites and the correct one should be almost complementary to 2L minus the fourth base.

Cassette boundaries in class 2 integrons can be identified in a similar way: TAATAAAATG is found adjacent to the

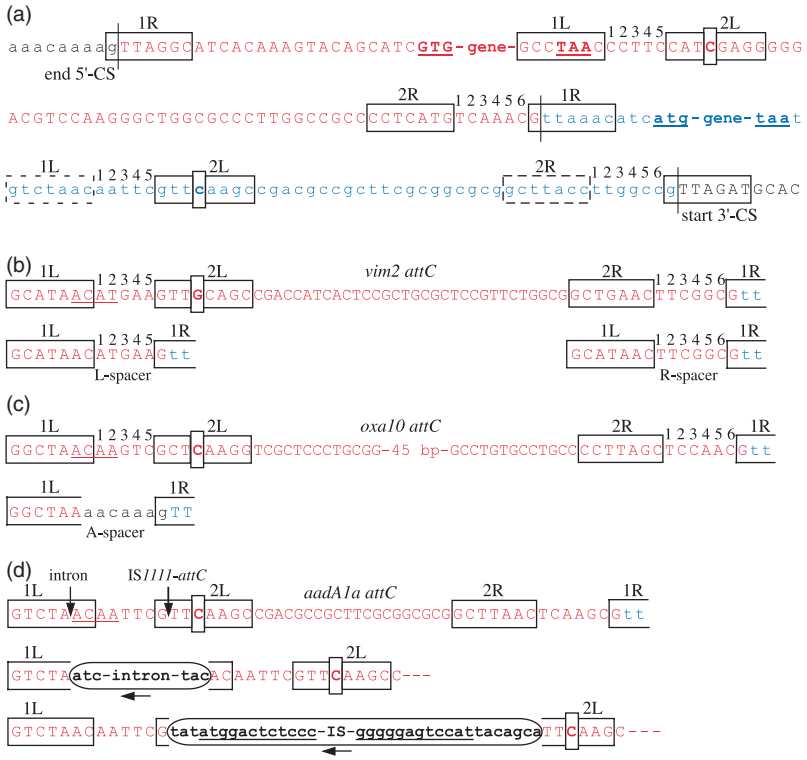

Fig. 5. Cassette boundaries and *attC* site modifications. (a) Identifying cassette boundaries and *attC* sites. Different cassettes are shown in red and blue with the 5′-CS and 3′-CS in black and alternating upper and lower case letters are also used to emphasize boundaries between different regions. The start and stop codons of cassette genes are shown (bold and underlined) with most of the gene sequence omitted. Core sites are boxed and labeled, with the extra nucleotide in the 2L core site indicated. Vertical lines indicate recombination sites/cassette boundaries. The final G residue of each cassette is the first nucleotide of the 1R site of that cassette, but for simplicity is included as part of the 1R site of the following cassette. (b) L-spacer and R-spacer truncations illustrated by the *vim2 attC* site. (c) A-spacer truncation illustrated by the *oxa10 attC* site. (d) Insertions into *attC* sites illustrated by the *aadA1a attC* site. Vertical arrows on the top line show the insertion points of class C-*attC* GII introns and IS*1111*-*attC* elements. The position of the 4 nt intron-binding site (IBS1) is underlined here and also in the *vim2* and *oxa10 attC* sites in parts (b) and (c). In the middle line, the intron is represented as an oval and the ends of the intron sequence are shown. In the bottom line the IS*1111*-*attC* element (ISUncu*1*) is represented as an oval, and the ends of the IS are shown, with the subterminal inverted repeats underlined. The horizontal arrows indicate the direction of the IEP gene in the intron and the transposase gene in the IS.

start of the first cassette and TTAGAG defines the start of the *ybeA* cassette usually found at the end of arrays.

Cassettes with a truncated *attC* site

Short versions of some cassettes with precise deletions in *attC* have been identified. The first six bases of the 1L core site are still present, but the remainder of the *attC* site is replaced by the L or R spacer sequence (shown as -L or -R; Fig. 5b) or by the last seven bases of the *attI1* site (‘*attI1* spacer’, shown as -A; Fig. 5c), probably due to recombination between the incorrect pair of core sites (Partridge *et al.*, 2000; Ramirez *et al.*, 2008). Few studies have examined movement of such cassettes, but testing for the loss of resistance phenotype suggested that *aadA10*-A was not excised from |*oxa10*|*aadB*|*aadA10*-A| (Partridge *et al.*, 2002b). PCR analysis of the |*aacA4*|*aadA1a*-A|*oxa9*| array indicated that *oxa9* alone was not excised and *aadA1a*-A was excised rarely while *aadA1a*-A and *oxa9* together were excised at higher levels (Ramirez *et al.*, 2008). Such truncated cassettes may be more likely to travel with the next cassette in the array or, if located adjacent to the 3′-CS, may be unlikely to be excised from an integron. Loss of most of *attC* may also allow increased expression of downstream cassette gene(s) from the Pc promoter.

For a few cassettes (*aacA33*-, *arr4*-R, *fosB*-A) only a truncated version is currently available and was used in the FDB. For other cassettes of this type, the appropriate spacer was added as a manual annotation that appeared in the cassette array analysis. Examples of *aadA1a*-A, *aadA1a*-R, *aadA2*-L, *aadA2*-R, *aadA5*-R, *aadA6*-A, *aadA10*-A, *aadA16*-

Table 5. Cassette arrays in class 1 integrons with class C-*attC* GII introns

Intron*	GenBank accession number	Position †	Cassette array ‡	Species §	Location¶
*S.ma*I2	AF453998.1	2504–534	#aadB-**intron**^{gcu8}|aadA6-A|aacA4|smr2|oxa10#	*S. marcescens*	Greece
*E.c.*I7‖	AY785243.1	2383–414	-**intron**^{aadA6}|aadA1a|	*E. coli*	Norway
*A.g.*I1**	AF369871.1	4803–2878	|vim2|aacA4|aadA1a-**intron**^{aadA1a}|	*Acinetobacter*	Korea
*S.ma*I3	AY884051.1	4858–2933	|aacC6|aacA7|vim2|aadA1a-**intron**^{aadA1a}|	*S. marcescens*	Korea?
	EF207718.1	5337–3412	|aacC6|aacA7|vim2|aadA1a-**intron**^{aadA1a}|	*P. aeruginosa*	Korea?
	AY065966.1	4399–2474	|aacA4|vim2-L|fosE-**intron**^{aadA1a}|	*P. putida*	Korea?
P.r. I1	AY887109.1	3623–1698	|oxa30:aacA4.4|vim2-**intron**^{aadA1a}|	*P. rettgeri*	Korea
	DQ153217.1	3964–2039	|oxa30:aacA4.4|vim2-**intron**^{aadA1a}|	*K. pneumoniae*	Korea
*Kl.pn.*I1	DQ153218.1	3956–2031	|oxa30:aacA4.4|vim2-**intron**^{aadA1a}|	*K. pneumoniae*	Korea
*P.ae.*I1 ††	AY029772.1	5441–3515	|oxa30:aacA4.1|vim2-fosE-vim2|aadA1a-**intron**^{aadA1a}| ‡‡	*P. aeruginosa*	Korea
*S.ma*FI §§	AY030343.1	3818–1893	|vim2|qacF-**intron**^{aadA1a}|	*S. marcescens*	Korea
*S.e.*I1i¶¶	AM932669.1	4331–2406	|dfrA21|oxa129|aadA1a-**intron**^{aadA1a}|	*S.* Bredeney	Brazil
*Kl.pn.*I2	AJ971342.1	2678–758	|dfrA14-**intron**^{dfrA14}|arr2|cmlA1|PCS-1|oxa10-A|aadA1a|‖‖	*K. pneumoniae*	France
*S.t.*I1	AY204504.1	100–2058	|aadB-**intron**^{cmlA1}|	*S.* Typhimurium	USA or China
*V.ch.*I1	EU116440.1	3745–1769	|dfrA12|gcuF|aadA2-**intron**^{aadA2}|	*V. cholerae*	China

*ORFs within introns are often annotated as orfii/ORFII and orfiii/ORFIII in the original GenBank entry.

†Positions are given with the intron oriented in the direction of the IEP gene.

‡Names as superscripts indicate the *attC* site most closely related to the partial *attC* sequence found after the intron. |cas#, partial cassette; |*cas*-A|, cassette with the 1L-*attI1* spacer structure in place of the complete *attC* site; |*cas*-L|, cassette with the 1L-L spacer structure in place of the complete *attC* site; |*cas1:cas2*| fused cassette

§Full species names are *Escherichia coli*, *Klebsiella pneumoniae*, *Providencia rettgeri*, *Pseudomonas aeruginosa*, *Pseudomonas putida*, *Serratia marcescens*, *Vibrio cholerae* and *S.* X indicates *Salmonella enterica* ssp. enterica serovar X.

¶?, location assumed from the address of the submitting authors as no publication is available.

‖*E.c.*I7 is 99% identical to *S.ma*.I2.

**All introns in this section are 100% identical to *A.g.*I1.

††*Pae*.I1 is 99% identical to *A.g.*I1.

‡‡The *fosE* cassette is inserted within the *attC* site of the *vim2* cassette, between the first and second positions of the 2L core site.

§§*S.ma*.F1 is 99% identical to *A.g.*I1.

¶¶*S.e.*I1 is 97% identical to *A.g.*I1.

‖‖This cassette array also has ISKpn*10* in *cmlA1* and ISKpn*3* in *oxa10*-A.

A, *catB3*-R, *ges1*-A, *oxa10*-A, *oxa10*-R, *oxa13*-A, *vim2*-L and *vim2*-R were identified here. In the case of *oxa10*, the truncated *oxa10*-A version was more common in GenBank than the complete cassette ($n = 19$ vs. 7).

Group II introns targeting 1L of *attC* sites

Group II introns (Lambowitz & Zimmerly, 2004; Toro *et al.*, 2007) belonging to bacterial class C are known to insert after potential Rho-independent terminators. Several have been identified within *attC* sites, inserted after the fifth base of 1L (Fig. 5d) with the gene for the intron-encoded protein (IEP) in the opposite orientation to cassette genes. Phylogenetic analysis of IEPs suggests that those targeting *attC* sites form a distinct clade, termed class C-*attC* GII introns (Quiroga *et al.*, 2008). Several were identified here (Table 5) including some that have already been published (Centrón & Roy, 2002; Dai & Zimmerly, 2002; Sunde, 2005; Michael *et al.*, 2008; Quiroga *et al.*, 2008).

Recent experiments (Quiroga *et al.*, 2008) demonstrated that *S.ma*.I2 is able to insert into several different *attC* sites (*aadA1*, *aacA1*:*gcuG*, *imp1*, *oxa10*, *sat2* and *dfrA1*). These *attC* all have putative intron-binding sites (IBS1, TTGT; IBS3, TAR) on the bottom strand, complementary to the proposed exon-binding site (EBS1, AACA; EBS3, A+N) in *S.ma*.I2 and also found in the introns listed in Table 5. The TTGT site overlaps 1L and the L-spacer (underlined in Fig. 5b–d) and is found in many *attC* sites. *S.ma*.I2 was not inserted into the *aacA4* (GGGT) or *gcuH* (AGGT) *attC* sites (Quiroga *et al.*, 2008). Insertion of *S.ma*.I2 required the secondary structure of the *attC* site in addition to these short RNA–DNA matches and a gene cassette with *S.ma*.I2 inserted in *attC* could still be excised (Quiroga *et al.*, 2008).

*A.g.*I1 is found in several related arrays, all from *P. aeruginosa* apparently isolated in Korea. In one array, a partial *vim2* cassette precedes *A.g.*I1 but the sequence after the intron matches the *aadA1a attC* site. This structure may result from recombination between introns inserted in different cassettes or intron-mediated deletions, but it not clear whether *A.g.*I1 could insert into the *vim2 attC* site, which has ATGT at the IBS1 position (Fig. 5b). Other introns were also automatically annotated as flanked by one partial cassette and the *attC* site of a different cassette (Table 5), but in most cases (except |*aadB-S.t.*I1^{cmlA1}|) the *attC* sites of the flanking cassettes are the same length and closely related, making it

harder to exclude variations in *attC* site sequences as an explanation.

IS*1111-attC* IS targeting 2L of *attC* sites

IS of the IS*1111-attC* group of the IS*1111*-like family insert after the first nucleotide of the 2L core site of *attC* (Fig. 5d) with the transposase in the opposite orientation to cassette genes (Post & Hall, 2008; Tetu & Holmes, 2008). IS of the IS*1111*-like family target specific sequences and are also unusual in that their inverted repeats are located a few bases inside the IS boundaries (Fig. 5d) (Partridge & Hall, 2003), often leading to incorrect annotation of their ends. Several different IS*1111-attC* were identified here (Table 6) and more details are available in Tetu & Holmes (2008) and/or Post & Hall (2008). In two arrays, the regions flanking the IS were annotated as belonging to different cassettes but, as with some introns described above, the *attC* sites of the two cassettes were the same length and closely related.

Insertion of an IS*1111-attC* element presumably reduces Pc-mediated transcription of downstream cassettes, but this may be overcome by the presence of an outward-facing promoter in the IS itself. ISPa*21* has a suitably situated promoter (TTGGCC–17 bp–TTTCAT) (Poirel *et al.*, 2005) and the same sequence is present in all ISPa*21* variants and in ISUnCu*1*, while TTGGCC–17 bp–CTTCAT is found at the equivalent position in ISKpn*4*.

Hybrid gene cassettes

Some gene cassettes appear to be 'hybrids' formed by homologous recombination between two closely related cassettes. The *vim* cassette in DQ143913, identified as a hybrid but named *vim12* (Pournaras *et al.*, 2005), was automatically annotated as a partial *vim1* cassette followed by a partial *vim2* cassette. CARB-6 in AF030945 may be a *blaP1*/*blaP7* hybrid and a cassette in AJ878850 could be a hybrid of a cassette designated *aacA36* here and *aacA4*.

The majority of hybrids are those formed between the closely related (89% identical) *aadA1a* and *aadA2* cassettes (Gestal *et al.*, 2005). There are several known variants of each of these cassettes (Partridge *et al.*, 2002a; Gestal *et al.*, 2005) and a number of different recombination crossover regions. Previously unrecognized hybrids were therefore identified by adding 'artificial hybrid' sequences, consisting of one quarter *aadA1*, three-quarters *aadA2* or half *aadA1a* and half *aadA2* etc. to the FDB. *aadA2/1* hybrids were most common (Supporting Information, Table S9), with 10 different crossover regions identified and are associated with a limited number of different arrays. Two different *aadA1/2* hybrids and three examples of the same *aadA2/1/2* hybrid were also identified.

Hybrid cassettes raise nomenclature issues, as some have been given distinct names (e.g. *aadA3*, *aadA8*, *aadA21*, *vim12*), but identifying and annotating them as hybrids may be more useful. In combination with context data, such information may reveal more about the role of homologous recombination in movement of parts of integrons and creation of different multi-resistance regions and plasmid structures.

Cassettes with atypical *attC* sites

As stated above, each cassette gene is generally associated with one particular *attC* site, but there are exceptions. A variant of the *aadA1* cassette (M95287.4, 3311–4166) known as *aadA1b* (Recchia & Hall, 1995b) has changes at the end of the *attC* site in 2R and the R-spacer (GCTTA**C**CT**TGGC** CG vs. GCTTA**A**CT**CAAG**CG) (Stokes & Hall, 1992). Our analysis identified other cassettes with an *attC* site that differed more substantially from the expected one. Cassettes named *aacA29a* and *aacA29b* carry almost identical genes (Poirel *et al.*, 2001) but the *attC* sites differ near 2R, with

Table 6. IS*1111-attC* elements

IS	GenBank accession number	Position*	IR_R and right end†	IR_L and left end†
ISPa*25*	DQ393782.1	2443–3852	tat**ATGGACTCCTCC**	**GGAGGAGTCCAT**tccatca
ISUnCu*1*	AM932676.1	1520–2899	tat**ATGGACTCTCCC**	**GGGGGAGTCCAT**tacagca
ISPa*21*a‡	AY920928.1	3576–4949	tat**ATGGACTCTCCC**	**GGGAGAGTCCAT**tacagcc
ISPa*21*b	AJ704863.3	9822–11160§		
ISPa*21*c	AY660529.1	2432–3806		
ISPa*21*d	AM296017	2742–4117		
ISKpn*4*a¶	EF408254.1	2347–3723	cat**ATGGACTCTCCC**	**GGGAGAGTCCAT**tacagcc
ISKpn*4*b	AM749812.1	2022–3399		

*The positions are for the $IR_R \rightarrow IR_L$ direction, as seen in cassette arrays.

†The 12-bp inverted repeats are shown in bold and uppercase, the termini of the IS in lowercase.

‡ISPa*21* variants are 91–96% identical and their IR and termini are 100% identical.

§The end of ISPa*21*b is missing in the only example available.

¶ISKpn*4*a and b are 97% identical and their IR and termini are 100% identical.

part of the *aacA29b* version matching the end of the *oxa20 attC*. Three related arrays in *P. aeruginosa* isolates from Taiwan (DQ393784, 2461–3099; EF138817; EU090799) each contain one or two copies of an *aacA4* gene associated with an *attC* site of the expected 72 nt but only *c.* 77% identical to the typical *attC* site and with two differences from the *gcu3 attC*. AF364344 (117–755) and AY139599 each include a gene > 98% identical to the exemplar *aadB* gene associated with a 109-nt *attC* site most closely related to the one typically found in the *aacC1* cassette, rather than the usual 60-nt *attC*. In EU851865 (172–699), an *aacC1* gene is associated with the *attC* site found usually found in the *aadA1a* cassette.

These variant cassettes all match the exemplar cassette from the start until part way through the *attC* site. They may have been created by recombination in the *attC* site, either by homologous recombination between related stretches in two different *attC* sites or by abnormal site-specific recombination events such as a second round of cleavage and transfer normally avoided during IntI1-mediated reactions (MacDonald *et al.*, 2006). In contrast, a cassette in EU723083 (76–697) includes a gene that is 100% identical to *aacA3* but the entire *attC* site matches the one typically found in the *aadA5* cassette and the region from 1R to the start codon also differs from usual *aacA3* cassette. This may provide an example of the same progenitor gene becoming associated with two distinct *attC* sites. Cassettes with atypical *attC* sites appear rare (although it is possible that our analysis missed some), but they have implications for cassette nomenclature and may be useful epidemiological markers.

Other cassette variants

A number of cassettes that apparently consist of the start of one cassette and the end of an unrelated cassette were also identified (Table 7). These 'fused' cassettes have presumably arisen by deletions with endpoints in adjacent cassettes (Recchia & Hall, 1995b).

Several cassette variants with internal tandem duplications were also found. A *dfrA1* cassette with a 90-bp duplication in the coding sequence flanked by 15-bp direct repeats was associated with reduced levels of trimethoprim resistance (AJ400733.1, 2-end) (Gibreel & Skold, 2000). The first *dfrB1* cassette identified (U36276.2, 573–1057) includes a 72-bp duplication compared with the version most common in GenBank (Levings *et al.*, 2006). A *vim1* cassette, often the *vim4* variant, with a 170-bp duplication that includes all but the final C of the 1L site of *attC* (AY152821.1, 958–2041) (Patzer *et al.*, 2004; Scoulica *et al.*, 2004) is found 10 times in GenBank. A *blaP1* cassette, which appears to have a 52-bp duplication resulting in a 5′ extension of the gene, was also identified here (AB126603.1, 77–1172). These variant cassettes were added to the FDB as separate features and we have designated them as *dfrB1*^{d72}, *dfrA1*^{d90}, *vim1*^{d170} and *blaP1*^{d52}.

Table 7. Examples of cassette fusions

Fusion	1st*	2nd†	Overlap‡	GenBank accession number	Start	End
aacC2:*aacA4*	462	26-end	–§	AF355189.1	1423	2449
oxa10:*aacA4*¶	32	37-end	–	AY219651.1	2184	2818
oxa30:*aacA4*.1	190	16-end	T	AF227505.1	98	910
oxa30:*aacA4*.2	87	17-end	A	AY103455.1	73	781
oxa30:*aacA4*.3	94	30-end	CAAC	AY136758.1	308	1007
oxa30:*aacA4*.4	94	24-end	C	AY887109.1	141	849
catB3:*aacA4*	46	9-end	C	DQ321671.1	81	761
catB8:*aacA4*	324	23–615‖	T	S49888.1	254	> 1169
qacG:*aacA4***	109	28-end	A	EF118171.1	1353	3165
catB2:*aacA38*-A††	94	?-1L	?	EF382672.1	112 481	111 841
oxa10:*aadA6*-A	822	15-1L	TAAC	EU358785.1	701	2326
aadA1:*aadB*	17	2-end	–	AM932676.1	98	704
aadA1:*dfrA1*	13	23-end	GA	AY339625.2	10 411	10 976

*The extent of the first cassette present, from position 1 to the position indicated.

†The extent of the second cassette present, where position 1 is start of the cassette.

‡Bases at the junction between the two cassette fragments that could be derived from either cassette.

§A T residue that does not appear to be derived from either cassette is present at the junction between the two cassette fragments.

¶The only example to date is found in a class 3 integron.

‖The sequence in S49888 ends within the cassette.

**ISAeca*1* is inserted in the *qacG* fragment at positions 1439–2529 in EF118171.1, flanked by a 2 nt direct duplication (AC).

††A complete version of the *aacA38* cassette has not yet be been identified; thus, the start position and any possible overlapping bases cannot be defined.

Enhancing transcription of cassette-borne genes, for example *oxa10*

Expression of most cassette genes relies on the integron-borne Pc promoter and is influenced by the position of the cassette in the array. While a few cassettes carry an internal promoter, the *oxa10* gene may be expressed from a promoter in an adjacent region. As indicated above, *oxa10*-A is more common than the complete *oxa10* cassette in GenBank and most copies are preceded by a 161-bp region (7331–7491 in AF205943.1) that may be the start of a cassette, but the complete version has never been identified. This region, designated PCS-1 here, contains two overlapping promoters (TTGAAG–17 bp–TAAAGT and TTTAAA–16 bp–TCTGAT) (Naas *et al.*, 2001) and association with PCS-1 may thus allow transcription of *oxa10* independently of Pc. |PCS-1| *oxa10*-A| has been found in a few related arrays, followed by either *aadA1a* or *aacA4*, suggesting limited movement, but has been seen in a number of species and geographic locations.

Variations in the *aacA4* cassette that may provide translational signals

A few cassettes do not appear to carry a suitably positioned RBS, and in these cases translation of the cassette gene may be enhanced by proximity to a short ORF (ORF11) within the *attI* site when the cassette is first in the array (Hanau-Berçot *et al.*, 2002). An example is *accA4*, which has a GTG start codon at positions 25–27 of the cassette (Hanau-Berçot *et al.*, 2002). In several sequences, the *aacA4* gene is fused to ORF11 as a result of changes in the *att1* site (Fig. 6a) and several other modifications to *aacA4* may also enhance expression. Many of the fused cassettes listed in Table 7 include *aacA4* as the second partial cassette, and in these cases the RBS and start codon of the first gene are presumably used. A truncated cassette with a spacer instead of a complete *attC* site preceding a complete *aacA4* cassette may also provide an RBS and/or start codon (Fig. 6b).

Many of these modifications predict short N-terminal extensions of AacA4 and the variation in observed sizes of AacA4 proteins (e.g. Casin *et al.*, 1998; Hanau-Berçot *et al.*, 2002) and available N-terminal sequences (Tran van Nhieu & Collatz, 1987; Dery *et al.*, 2003) confirms some of these. In some cases, for example |*aadA6*-A|*aacA4*| in AF453998 (Centrón & Roy, 2002) and |*aacC2*:*aacA4*| in AF355189 (Dubois *et al.*, 2002), long ORFs are created by fusion of almost full-length genes.

A distinctive region associated with an AacA4 protein variant

In addition to the modifications to *aacA4* described above, several point mutations are known to result in differences in the aminoglycoside resistance phenotype conferred. A change (T to C) at position 329 of the *aacA4* cassette results in Leu102Ser (numbered from the GTG start codon) in the

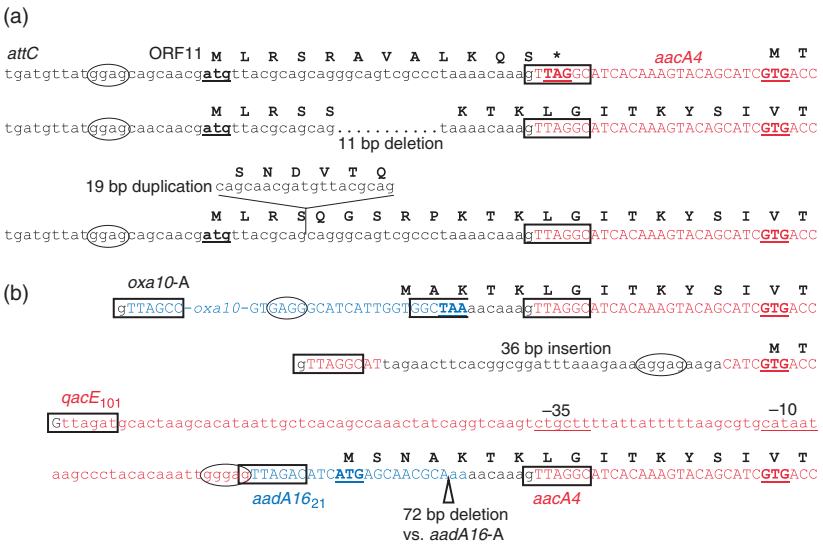

Fig. 6. Variations in the start of the *aacA4* gene cassette. Cassette sequences are shown in red and blue and other regions in black, with alternating upper and lower case letters used to emphasize boundaries between different regions. Potential RBSs are indicated by ovals and -35 and -10 regions are underlined and labeled. Start and stop codons are in bold and underlined and amino acid sequences are shown above. Core sites are boxed. (a) Alterations in the *attI1* site. Top, standard *attI1* showing ORF11; middle, an 11-bp deletion in *attI1* (six in GenBank, e.g. AJ621187); bottom, a 19-bp duplication in *attI1* (30 in GenBank, e.g. AB212941). (b) Modifications associated with *aacA4cr* variants. Top, *aacA4* preceded by *oxa10*-A; middle, *aacA4* cassette with a 36-bp insertion replacing positions 9–20; bottom, *aacA4* preceded by the $qac_{101}aadA16_{21}$-A structure. The two lowercase residues in blue (aa) are presumably derived from the 1L site in *aadA16*-A.

AacA4 protein. Both variants confer resistance to tobramycin, but additional resistance to amikacin correlates with T329/Leu102 [an *aac*(6′)-*I* phenotype, *aacA4*Ak here] and resistance to gentamicin with C329/Ser102 [an *aac*(6′)-*II* phenotype; *aacA4*Gm here] (Rather *et al.*, 1992). A variant with Gln101Leu as well as Leu102Ser confers significant resistance to both amikacin and gentamicin (Casin *et al.*, 2003). The more recently identified AAC(6′)-Ib-cr variant is less effective against aminoglycosides, but confers additional low-level resistance to fluoroquinolones (Robicsek *et al.*, 2006). All currently known cassettes encoding this variant have T329/Leu102, the same mutation at position 514 (GAT to TAT; Asp164Tyr) and synonymous mutations at position 283, TGG to AGG (designated *aacA4*crA here) or to CGG (*aacA4*crC) giving Trp87Arg.

Preliminary examination of the distribution of different *aacA4* variants suggests that *aacA4*Ak and *aacA4*Gm are each found in a variety of arrays. However, the majority of *aacA4*cr cassettes are preceded by a distinctive structure consisting of first 101 bp of the *qacE* cassette (or the 3′-CS, both are identical in this region; $qacE_{101}$ here), the first 21 nt of the *aadA16* cassette ($aadA16_{21}$ here) and the sequence AAAACAAAG (Fig. 6b). $qacE_{101}$ is also found preceding most examples of *aacA29* and complete *aadA16* cassettes identified to date and may have resulted from site-specific recombination at a secondary site (GATATA, where G is position 101) in the *qacE* cassette/3′-CS (Poirel *et al.*, 2001). The remainder of this structure may have been created by deletion of 792 bp from $|qacE_{101}|aadA16$-A|, which is found in AY740681 (Table 8). $qacE_{101}$ includes the weak *qacE* promoter (Guerineau *et al.*, 1990) and a potential RBS and translation from the *aadA16* start codon would give an AacA4 protein with a 15 amino acid N-terminal extension (Fig. 6b) (Robicsek *et al.*, 2006).

In EF636461 positions 9–20 of the *aacA4* cassette have been replaced by a 36-bp sequence that provides a potential RBS (Fig. 6b), but similarity of the downstream cassettes (Table 8) to those found in arrays with the $|qacE_{101}|aadA16_{21}$-A structure suggests homologous recombination within the cassette array. The remaining *aacA4*cr cassette (EU161636) is preceded by *oxa10*-A (Fig. 6b). Determining the context of other examples of *aacA4*cr may reveal more about the spread of this important cassette variant.

Using silent mutations to track epidemiology, for example *ges* cassettes

Characteristic large-scale modifications to cassettes are clear epidemiological markers. However, point mutations, both silent and those resulting in amino acid changes that may or may not affect resistance phenotype, are also potentially useful. As part of developing automated methods to identify cassette variants, we found some interesting patterns in the distribution of variants of the *ges1* cassette.

ges1 confers resistance to penicillins and some cephalosporins, but not aztreonam, β-lactamase inhibitors, or cephamycins (Poirel *et al.*, 2000). Other *ges* genes are minor variants encoding proteins with 1–3 amino acid changes (Table 9). GES-2, GES-4, GES-5 and GES-6 (all Gly170Asn or Ser) have appreciable carbapenemase activity, particularly GES-5 (Smith *et al.*, 2007), while GES-9 (Gly243Ser) has increased activity against aztreonam (Poirel *et al.*, 2005).

ges cassettes have been found in several unrelated arrays, often as the first cassette and sometimes in truncated form as *ges*-A (Table 9). Detailed sequence analysis suggested that *ges* cassettes from different geographic locations (Greece, Japan, China; Table 9) have characteristic combinations of silent mutations, with mutations that result in phenotypic changes superimposed. This provides evidence of local evolution and suggests that the mutations that affect phenotype, particularly to give *ges5*, may have occurred on separate occasions.

It is important to distinguish variants that have arisen locally from those acquired from other regions, as their associations with other cassettes and resistance genes may differ. Moreover, a sudden predominance of 'migrant' genes may signal greater transmissibility risk than a local variant that has arisen by point mutation in the presence of strong selection pressure.

Trying to understand cassette epidemiology

A better understanding of how different cassettes move within the gene pool would greatly assist in tracking the spread of antibiotic resistance genes, but trying to find meaningful patterns in currently available information is difficult. The data available in GenBank and analysed here is unlikely to be representative of natural bacterial populations due to biases in, for example, the selection of isolates studied and of sequences deposited. However, these data are the most easily accessible and a limited analysis, with reference to complementary published data, may provide useful information to direct more systematic studies. The numbers of complete copies of each cassette or *gcu* identified in sequences in GenBank were therefore compiled (Tables 1 and 2) and the distribution patterns of selected cassettes examined.

Nearly half of the resistance cassettes (60/132) and most (49/62) *gcu* were found in ≤ 2 GenBank entries and some of these cassettes may indeed be uncommon. The *aacA2* [*aac(6′)-Id*] gene was rare in early surveys of aminoglycoside-resistant strains (Shaw *et al.*, 1993) and is found in only one sequence in GenBank (X12618, apparently deposited in

Table 8. Cassette arrays that include *aacA4*cr variants and related cassette arrays

GenBank accession number	Cassette array*	Species†	Location‡
AY509609	|PCS-1|oxa10-A|**aacA4Ak** |vim1^{d170}|	*P. aeruginosa*	Hungary
DQ984668	|PCS-1|oxa10-A|**aacA4Gm** |cmlA1|	*P. aeruginosa*	Greece?
EU161636	|PCS-1|oxa10-A|**aacA4crA**|cmlA1|	*P. aeruginosa*	Hungary
EF522838	|vim2|PCS-1|oxa10-A|**aacA4Ak** |qacK|aadA1a|	*P. putida*	Singapore
AY560837	|vim2|36-**aacA4Gm** |blaP1|aadA2|	*P. aeruginosa*	Portugal
EU052800	|36-**aacA4Gm** |aadA1a|	*E. cloacae*	Argentina
EF636461	|36-**aacA4crA**|oxa30|catB3|arr2|	*K. pneumoniae*	Argentina
AY740681	|qacE$_{101}$|aadA16-A |**aacA4Gm** |oxa30|catB3|§	*A. punctata*	France?
AJ971343	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crA**|oxa30|catB3|arr2|	*K. pneumoniae*	France
AY259086	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|oxa30|catB3|arr2|	*E. coli*	China
EF415651	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|oxa30|catB3|arr2|	*E. coli*	Singapore?
AY458016	IS*26*#qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|oxa30|catB3#IS*26*¶	*E. coli*	Canada
EU495237	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|oxa30|arr2|dfrA27|qacE$_{101}$|aadA16|‖	*K. pneumoniae*	France
EU495238	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|oxa30|arr2|dfrA27|qacE$_{101}$|aadA16|‖	*K. ascorbata*	France
EU675686	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|arr2|dfrA27|qacE$_{101}$|aadA16|**	*E. coli*	China
EU543272	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crA**|arr2|	*K. pneumoniae*	China
EU678897	|arr2|dfrA27|qacE$_{101}$:aadA16|	*V. cholerae*	China?
EU195449	|qacE$_{101}$|aadA16$_{21}$-A|**aacA4crC**|	*K. pneumoniae*	China
DQ303918	#aadA16$_{21}$-A|**aacA4crC**#	*E. coli*	China
EF100892	#aadA16$_{21}$-A|**aacA4crC**#	*K. pneumoniae*	Slovenia

*The crA variant has AGG encoding Arg87, while crC has CGG encoding Arg87. |*oxa10*-A| has the 1L-*attI1* spacer structure in place of the complete *attC* site. |36-*aacA4*|, a cassette with a 36-bp insertion replacing positions 9–20 of the standard *aacA4* cassette (Fig. 6b).

†Full species names are *Aeromonas punctata*, *Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumoniae*, *Kluyvera ascorbata*, *Pseudomonas aeruginosa*, *Vibrio cholerae*.

‡?, location assumed from the address of the submitting authors as no publication is available.

§Most of the *oxa30* cassette (nt 8-919/1004) has apparently been replaced by a region containing a gene 78% identical to a Txe/YoeB family protein in GenBank accession number CP000529.

¶In this sequence from plasmid pC15-1a, two copies of IS*26* flank the cassette array and the 5′-CS and 3′-CS are not present.

‖IS*26* is inserted in the 5′-CS at a different position in each sequence and a 217-bp deletion encompasses the start of the *intI1* gene but not Pc.

**There are two additional changes in this AacA4crC protein, Ser4Cys and Glu23Gly, not found in any other AacA4 variants.

1988). In contrast, *aacA4* and *aadB* were common in these surveys (Shaw *et al.*, 1993) and are also common in sequences in GenBank (Table 1). *vim2* was the most common metallo-β-lactamase (MBL) cassette in GenBank and VIM-2 appears to be the most widespread MBL worldwide, at least in *P. aeruginosa* (Walsh, 2008). A few other cassettes were identified frequently in GenBank (Table 1) and it seems reasonable to conclude that at least some of these are more widespread than the majority of cassettes that are found in only a few sequences.

A simple analysis of the distribution patterns of complete examples of these apparently 'successful' cassettes in arrays are in class 1 integrons also revealed some potentially interesting differences, suggesting that different cassettes may spread in different ways. *aacA4* was found in many different arrays (*c.* 90) with none dominating, but was generally at the first or second position (*c.* 40% for each). *aadB* and *vim2* were each found in a variety of arrays (*c.* 35) and as the first cassette in most (*c.* 70% and *c.* 60%, respectively). *aadA1a* was commonly found as a lone cassette (*c.* 20%), in |*dfrA1*|*aadA1a*| (*c.* 25%) or the last cassette (in 54 of 56 other arrays). *dfrA1* was largely restricted to two arrays, |*dfrA1*|*aadA1*| (*c.* 50%) and |*dfrA1*|*gcuC*| (*c.* 30%). The |*dfrA12*|*gcuF*|*aadA2*| array accounted for almost all examples of both *dfrA12* and *gcuF* but only *c.* 30% of *aadA2*, which was also commonly seen as a lone cassette (*c.* 40%). Most examples of *aadA5* (*c.* 80%) and all examples of *dfrA17* were found in |*dfrA17*|*aadA5*|. Many of these apparently common arrays also dominate in surveys from which no sequences were deposited in GenBank (e.g. Yu *et al.*, 2003; Machado *et al.*, 2007).

Factors influencing cassette epidemiology

Many different factors are likely to contribute to the epidemiology of a gene cassette, one of which is *attC* site activity. However, relatively few *attC* sites have actually been

Table 9. Comparison of *ges* gene variants, GES proteins and cassette arrays carrying *ges* cassettes by geographic location

Location	Species	Nucleotide changes*	Cassette array†	Amino acid changes‡	GenBank accession numbers§
Portugal	*K. pneumoniae*		|**ges1**|oxa10:aacA4|¶		AY219651
France	*P. aeruginosa*		|aacC2:aacA4|**ges1**|		AF355189
France	*P. aeruginosa*	T601C	|**ges9**|aacA4|gcuD|aadB|‖	G243S	**AY920928**
Fr. Guiana**	*K. pneumoniae*	C634T	|**ges1-A**|aacA4|dfrA15|cmlA1|aadA2-R|		**AF156486**
Brazil	*P. aeruginosa*		|gcu14|**ges1-A**|aacA4#		DQ236170
Brazil	*P. aeruginosa*		|**ges5-A**|aacA4#	G170S	DQ236171
Brazil	*K. pneumoniae*	G864GG	partial **ges7** gene	E104K	EF219163
Brazil	*K. pneumoniae*	G76C††	partial **ges7** gene	E104K	EF219164
Brazil	*P. aeruginosa*	not available‡‡	|**ges1**|**catB8**|		
S. Africa	*P. aeruginosa*		**ges2** gene only	G170N	**AF326355**
S. Africa	*P. aeruginosa*		|**ges2**|oxa5|aacC4|	G170N	AF347074
S. Africa	*P. aeruginosa*		|**ges5**#	G170S	EF190326
S. Africa	*P. aeruginosa*		**ges5-like** gene only	A19E G170S	EF202187
Korea?	*K. pneumoniae*		|**ges5**|aacA3|oxa10|§§	G170S	DQ902344
China?	*P. aeruginosa*		partial **ges1** gene		DQ333893
Greece?	*P. aeruginosa*	G97A	|**ges1**|		EU598463
Greece	*P. aeruginosa*	G97A	|**ges8**|	A125L	**AF329699**
Greece	*E. coli*	G97A	**ges5** gene only	G170S	**AY494717**
Greece	*E. coli*	G97A	|aacA4|**ges7**|smr2|dfrA1|aadA1a|	E104K	AY260546
Greece	*E. cloacae*	G97A	**ges7** gene only	E104K	**AF208529**
Greece	*K. pneumoniae*	G97A	**ges6** gene only	E104K G170S	**AY494718**
Japan	*K. pneumoniae*	GA37-8AG T220C C1019T¶¶	|**ges3**|aacA1:gcuG|gcu6|	M61T E104K	**AB113580**
Japan	*K. pneumoniae*	GA37-8AG T220C C1019T	|**ges4**|aacA1:gcuG|gcu6|‖‖	M61T E104K G170S	AB116723
Japan	*K. pneumoniae*	GA37-8AG T220C C1019T	|**ges4**|aacA1:gcuG|‖‖	M61T E104K G170S	**AB116260**
China	*P. aeruginosa*	C202A G520C G598T	**partial ges5** gene	G170S	AY953375
China	*P. aeruginosa*	C202A G520C G598T	**partial ges5** gene	G170S	DQ660416
China	*P. aeruginosa*	C202A G520C G598T	**partial ges5** gene	G170S	DQ660417

*Silent nucleotide changes vs. the *ges1* cassette in AF355189 in addition to those resulting in amino acid changes.

†GES-3 and GES-4 have been used for two different pairs of enzymes (Lee & Jeong, 2005). The system used by http://www.lahey.org/Studies/(Weldhagen *et al.*, 2006; Walther-Rasmussen & Hoiby, 2007) has been used here. Partial *ges* genes/cassettes were not automatically annotated, but have been included in the analysis for completeness.

‡Positions are according to the ABL numbering scheme for class A β-lactamases (Ambler *et al.*, 1991).

§GenBank entries listed at http://www.lahey.org/Studies/are in bold.

¶This array is found in a class 3 integron.

‖Two copies of ISPa*21* are found in this array, in the *ges9* and *aadB attC* sites.

**French Guiana, situated in South America.

††This sequence also includes deletions/insertions (T856△ G873GC T874△) that would cause amino acid changes, but these may be errors.

‡‡There is no GenBank entry for this sequence. The information is from Castanheira *et al.* (2004).

§§ISPa25 is inserted in *oxa10*.

¶¶This mutation is the only one in the *attC* site.

‖‖Obtained from the same isolate.

tested for activity and the variety of methods used makes it difficult to compare results across studies and to relate *attC* site structure to activity. Several of the most common cassettes in GenBank (*aadA1a*, *aadA2*, *aadB*) have 'classical' 60-nt *attC* sites, the type that is probably the best studied and may be among the most active. However, this type of *attC* site is also found in many other cassettes that are rarer in GenBank. The *aadA1b* variant has a mismatch in the 2L/2R pair at a position that appears to be important in integrase binding and was much less common in GenBank than *aadA1a* ($n = 13$ vs. 259). The *dfrA1 attC* site recombined only at an extremely low frequency (Biskri *et al.*, 2005), which may help to explain why this cassette, although common, appears largely restricted to only two different arrays. It is also interesting that the *veb1* cassette, which has an *attC* site that appears atypical compared with other *attC* found in MRI, may have limited mobility: available examples of the complete cassette are always found preceding *aadB* in related arrays and a truncated *veb* cassette has been found flanked by short repeated elements, which may provide an alternative means of movement (Zong *et al.*, 2009).

The epidemiology of a cassette is also likely to be influenced by the phenotype it confers and how strongly this phenotype is selected. The phenotype will depend on the intrinsic activity of the encoded protein and the amount produced. Expression of the cassette gene is influenced by the strength of the associated Pc promoter, the position of the cassette in the array, the presence of any additional promoters and whether the cassette carries an RBS or is positioned to take advantage of an adjacent one. Of the cassettes discussed above, those conferring resistance to antibiotics that now have little clinical importance (*aadA*, streptomycin and spectinomycin resistance) were commonly found as the last cassette in an array, where they might be expected to be poorly expressed. In contrast, cassettes conferring resistance to antibiotics that are clinically relevant (*aadB*, gentamicin; *aacA4*, gentamicin or amikacin; *vim2*, β-lactams and potentially carbapenems) appeared most often in the first or second position, which is more favourable for expression. Many arrays in GenBank include both aminoglycoside and β-lactam resistance cassettes, which may reflect coselection, as well as a particular interest (and therefore surveillance/reporting bias) in these phenotypes.

The mobility of integrons themselves and of transposons and larger multiresistance regions carrying integrons will also influence the distribution of different gene cassettes. Selection pressure on linked resistance genes associated with other mobile elements (e.g. IS*CR1*) will also play a part, as will the copy number, transferability and host range of plasmid vehicles. The association of class 1 integrons with Tn*21*-like transposons is obviously likely to be important. The *aadA1a* cassette may owe much of its apparent success to an early association with Tn*21*, derivatives of which have disseminated widely (Liebert *et al.*, 1999). An association with a particularly successful larger mobile structure(s) or plasmid(s) may also help to explain why arrays such as |*dfrA12*|*gcuF*|*aadA2*| and |*dfrA17*|*aadA5*| appear common. However, as only about 10% of class 1 integron sequences examined here include any sequence beyond a complete 5′-CS or 3′-CS, little can be inferred about the contributions of these different factors at present.

Concluding remarks

Our novel automated approach (G. Tsafnat *et al.*, unpublished data) enabled integration of large amounts of data that would have been very difficult and extremely time consuming to handle manually and allowed identification of cassettes based on their context. Future analyses could potentially be modified to incorporate automated text searching (Grivell, 2002) to extract and include data only currently available in published papers. It may also be possible to use this information to more easily distinguish isolates from different sources (e.g. clinical, animal, environmental), geographical regions and years.

It is clear that useful information in available sequences may be overlooked because cassette genes, rather than complete cassettes, are identified and annotated. Many different modifications to cassettes were revealed, some of which may act as useful epidemiological signals for tracking spread of cassettes and evolution of cassette arrays. Cassette sequences with characteristic minor variations may also provide extra information about the epidemiology of cassettes and the evolution of arrays and we are developing additional grammar rules to distinguish these variants.

Our analysis further suggests that new resistance gene cassettes, including those conferring resistance to additional classes of antibiotics, continue to be acquired by MRI. Any attempts to predict the epidemic potential of these emerging cassettes clearly require a much better understanding of the relative influences of all the factors contributing to the spread of a gene cassette and how much this varies between different cassettes. While this analysis of MRI sequences in GenBank may provide a few hints, more systematic experimental exploration of these factors, particularly the wider genetic contexts of cassette arrays, is needed.

Acknowledgements

S.R.P. and G.T. are supported by separate NSW Health Capacity Building and Infrastructure Grants to CIDM and CHI.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Ambler RP, Coulson AF, Frere JM, Ghuysen JM, Joris B, Forsman M, Levesque RC, Tiraby G & Waley SG (1991) A standard numbering scheme for the class A β-lactamases. *Biochem J* **276**: 269–270.

Arakawa Y, Murakami M, Suzuki K, Ito H, Wacharotayankun R, Ohsuka S, Kato N & Ohta M (1995) A novel integron-like element carrying the metallo-β-lactamase gene bla_{IMP}. *Antimicrob Agents Ch* **39**: 1612–1615.

Baquero F (2004) From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nat Rev Microbiol* **2**: 510–518.

Barker A, Clark CA & Manning PA (1994) Identification of VCR, a repeated sequence associated with a locus encoding a hemagglutinin in *Vibrio cholerae* O1. *J Bacteriol* **176**: 5450–5458.

Barlow RS & Gobius KS (2006) Diverse class 2 integrons in bacteria from beef cattle sources. *J Antimicrob Chemoth* **58**: 1133–1138.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW (2009) GenBank. *Nucleic Acids Res* **37**: D26–D31.

Biskri L & Mazel D (2003) Erythromycin esterase gene *ere*(A) is located in a functional gene cassette in an unusual class 2 integron. *Antimicrob Agents Ch* **47**: 3326–3331.

Biskri L, Bouvier M, Guerout AM, Boisnard S & Mazel D (2005) Comparative study of class 1 integron and *Vibrio cholerae* superintegron integrase activities. *J Bacteriol* **187**: 1740–1750.

Bissonnette L, Champetier S, Buisson JP & Roy PH (1991) Characterization of the nonenzymatic chloramphenicol resistance (*cmlA*) gene of the In4 integron of Tn*1696*: similarity of the product to transmembrane transport proteins. *J Bacteriol* **173**: 4493–4502.

Boucher Y, Labbate M, Koenig JE & Stokes HW (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol* **15**: 301–309.

Bouvier M, Demarre G & Mazel D (2005) Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J* **24**: 4356–4367.

Boyd DA & Mulvey MR (2006) OXA-1 is OXA-30 is OXA-1. *J Antimicrob Chemoth* **58**: 224–225.

Brízio A, Conceição T, Pimentel M, Da Silva G & Duarte A (2006) High-level expression of IMP-5 carbapenemase owing to point mutation in the -35 promoter region of class 1 integron among *Pseudomonas aeruginosa* clinical isolates. *Int J Antimicrob Ag* **27**: 27–31.

Bunny KL, Hall RM & Stokes HW (1995) New mobile gene cassettes containing an aminoglycoside resistance gene, *aacA7*, and a chloramphenicol resistance gene, *catB3*, in an integron in pBWH301. *Antimicrob Agents Ch* **39**: 686–693.

Cameron FH, Groot Obbink DJ, Ackerman VP & Hall RM (1986) Nucleotide sequence of the AAD(2″) aminoglycoside adenylyltransferase determinant *aadB*. Evolutionary relationship of this region with those surrounding *aadA* in R538-1 and *dhfrII* in R388. *Nucleic Acids Res* **14**: 8625–8635.

Casin I, Bordon F, Bertin P, Coutrot A, Podglajen I, Brasseur R & Collatz E (1998) Aminoglycoside 6′-*N*-acetyltransferase variants of the Ib type with altered substrate profile in clinical isolates of *Enterobacter cloacae* and *Citrobacter freundii*. *Antimicrob Agents Ch* **42**: 209–215.

Casin I, Hanau-Bercot B, Podglajen I, Vahaboglu H & Collatz E (2003) *Salmonella enterica* serovar Typhimurium bla_{PER-1} carrying plasmid pSTI1 encodes an extended-spectrum aminoglycoside 6′-*N*-acetyltransferase of type Ib. *Antimicrob Agents Ch* **47**: 697–703.

Castanheira M, Mendes RE, Walsh TR, Gales AC & Jones RN (2004) Emergence of the extended-spectrum β-lactamase GES-1 in a *Pseudomonas aeruginosa* strain from Brazil: report from the SENTRY antimicrobial surveillance program. *Antimicrob Agents Ch* **48**: 2344–2345.

Centrón D & Roy PH (2002) Presence of a group II intron in a multiresistant *Serratia marcescens* strain that harbors three integrons and a novel gene fusion. *Antimicrob Agents Ch* **46**: 1402–1409.

Clark NC, Olsvik O, Swenson JM, Spiegel CA & Tenover FC (1999) Detection of a streptomycin/spectinomycin adenylyltransferase gene (*aadA*) in *Enterococcus faecalis*. *Antimicrob Agents Ch* **43**: 157–160.

Collis CM & Hall RM (1992) Gene cassettes from the insert region of integrons are excised as covalently closed circles. *Mol Microbiol* **6**: 2875–2885.

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

Collis CM, Grammaticopoulos G, Briton J, Stokes HW & Hall RM (1993) Site-specific insertion of gene cassettes into integrons. *Mol Microbiol* **9**: 41–52.

Collis CM, Kim MJ, Stokes HW & Hall RM (1998) Binding of the purified integron DNA integrase Intl1 to integron- and cassette-associated recombination sites. *Mol Microbiol* **29**: 477–490.

Collis CM, Recchia GD, Kim MJ, Stokes HW & Hall RM (2001) Efficiency of recombination reactions catalyzed by class 1 integron integrase IntI1. *J Bacteriol* **183**: 2535–2542.

Collis CM, Kim MJ, Partridge SR, Stokes HW & Hall RM (2002) Characterization of the class 3 integron and the site-specific recombination system it determines. *J Bacteriol* **184**: 3017–3026.

Correia M, Boavida F, Grosso F, Salgado MJ, Lito LM, Cristino JM, Mendo S & Duarte A (2003) Molecular characterization of a new class 3 integron in *Klebsiella pneumoniae*. *Antimicrob Agents Ch* **47**: 2838–2843.

Dai L & Zimmerly S (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* **30**: 1091–1102.

Da Silva GJ, Correia M, Vital C, Ribeiro G, Sousa JC, Leitao R, Peixe L & Duarte A (2002) Molecular characterization of

bla$_{IMP-5}$, a new integron-borne metallo-β-lactamase gene from an *Acinetobacter baumannii* nosocomial isolate in Portugal. *FEMS Microbiol Lett* **215**: 33–39.

Demarre G, Frumerie C, Gopaul DN & Mazel D (2007) Identification of key structural determinants of the IntI1 integron integrase that influence *attC* x *attI1* recombination efficiency. *Nucleic Acids Res* **35**: 6475–6489.

Dery KJ, Soballe B, Witherspoon MS, Bui D, Koch R, Sherratt DJ & Tolmasky ME (2003) The aminoglycoside 6′-*N*-acetyltransferase type Ib encoded by Tn*1331* is evenly distributed within the cell's cytoplasm. *Antimicrob Agents Ch* **47**: 2897–2902.

Dubois V, Poirel L, Marie C, Arpin C, Nordmann P & Quentin C (2002) Molecular characterization of a novel class 1 integron containing *bla*$_{GES-1}$ and a fused product of *aac3-Ib*/*aac*(6′)-*Ib*′ gene cassettes in *Pseudomonas aeruginosa*. *Antimicrob Agents Ch* **46**: 638–645.

Elbourne LD & Hall RM (2006) Gene cassette encoding a 3-*N*-aminoglycoside acetyltransferase in a chromosomal integron. *Antimicrob Agents Ch* **50**: 2270–2271.

Fluit AC & Schmitz FJ (1999) Class 1 integrons, gene cassettes, mobility, and epidemiology. *Eur J Clin Microbiol* **18**: 761–770.

Fluit AC & Schmitz FJ (2004) Resistance integrons and super-integrons. *Clin Microbiol Infec* **10**: 272–288.

Fonseca EL, Dos Santos Freitas F, Vieira VV & Vicente AC (2008) New *qnr* gene cassettes associated with superintegron repeats in *Vibrio cholerae* O1. *Emerg Infect Dis* **14**: 1129–1131.

Francia MV, Zabala JC, de la Cruz F & Garcia Lobo JM (1999) The IntI1 integron integrase preferentially binds single-stranded DNA of the *attC* site. *J Bacteriol* **181**: 6844–6849.

Gassama Sow A, Diallo MH, Gatet M, Denis F, Aidara-Kane A & Ploy MC (2008) Description of an unusual class 2 integron in *Shigella sonnei* isolates in Senegal (sub-Saharan Africa). *J Antimicrob Chemoth* **62**: 843–844.

Gestal AM, Stokes HW, Partridge SR & Hall RM (2005) Recombination between the *dfrA12*-orfF-*aadA2* cassette array and an *aadA1* gene cassette creates a hybrid cassette, *aadA8b*. *Antimicrob Agents Ch* **49**: 4771–4774.

Gibreel A & Skold O (2000) An integron cassette carrying *dfr1* with 90-bp repeat sequences located on the chromosome of trimethoprim-resistant isolates of *Campylobacter jejuni*. *Microb Drug Resist* **6**: 91–98.

Gillings M, Boucher Y, Labbate M, Holmes A, Krishnan S, Holley M & Stokes HW (2008) The evolution of class 1 integrons and the rise of antibiotic resistance. *J Bacteriol* **190**: 5095–5100.

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

Grindley ND, Whiteson KL & Rice PA (2006) Mechanisms of site-specific recombination. *Annu Rev Biochem* **75**: 567–605.

Grivell L (2002) Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO Rep* **3**: 200–203.

Grune D & Jacobs C (2007) *Parsing Techniques: A Practical Guide*. Springer, New York.

Guerineau F, Brooks L & Mullineaux P (1990) Expression of the sulfonamide resistance gene from plasmid R46. *Plasmid* **23**: 35–41.

Hall RM & Collis CM (1995) Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Microbiol* **15**: 593–600.

Hall RM, Brookes DE & Stokes HW (1991) Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point. *Mol Microbiol* **5**: 1941–1959.

Hanau-Berçot B, Podglajen I, Casin I & Collatz E (2002) An intrinsic control element for translational initiation in class 1 integrons. *Mol Microbiol* **44**: 119–130.

Hansson K, Sköld O & Sundström L (1997) Non-palindromic *attI* sites of integrons are capable of site-specific recombination with one another and with secondary targets. *Mol Microbiol* **26**: 441–453.

Hansson K, Sundström L, Pelletier A & Roy PH (2002) IntI2 integron integrase in Tn*7*. *J Bacteriol* **184**: 1712–1721.

Heir E, Lindstedt BA, Leegaard TM, Gjernes E & Kapperud G (2004) Prevalence and characterization of integrons in blood culture *Enterobacteriaceae* and gastrointestinal *Escherichia coli* in Norway and reporting of a novel class 1 integron-located lincosamide resistance gene. *Ann Clin Microbiol Antimicrob* **3**: 12–20.

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

Johansson C, Kamali-Moghaddam M & Sundström L (2004) Integron integrase binds to bulged hairpin DNA. *Nucleic Acids Res* **32**: 4033–4043.

Kamali-Moghaddam M & Sundstrom L (2000) Transposon targeting determined by resolvase. *FEMS Microbiol Lett* **186**: 55–59.

Labbate M, Roy Chowdhury P & Stokes HW (2008) A class 1 integron present in a human commensal has a hybrid transposition module compared to Tn*402*: evidence of interaction with mobile DNA from natural environments. *J Bacteriol* **190**: 5318–5327.

Lambowitz AM & Zimmerly S (2004) Mobile group II introns. *Annu Rev Genet* **38**: 1–35.

Lee SH & Jeong SH (2005) Nomenclature of GES-type extended-spectrum β-lactamases. *Antimicrob Agents Ch* **49**: 2148; author reply 2148–2150.

Leung S, Mellish C & Robertson D (2001) Basic gene grammars and DNA-ChartParser for language processing of *Escherichia coli* promoter DNA sequences. *Bioinformatics* **17**: 226–236.

Lévesque C & Roy P (1993) PCR analysis of integrons. *Diagnostic Molecular Microbiology: Principles and Applications* (Persing DH, Smith TF, Tenover FC & White TJ, eds), pp. 590–594. American Society for Microbiology, Washington, DC.

Lévesque C, Brassard S, Lapointe J & Roy PH (1994) Diversity and relative strength of tandem promoters for the antibiotic-resistance genes of several integrons. *Gene* **142**: 49–54.

Levings RS, Partridge SR, Lightfoot D, Hall RM & Djordjevic SP (2005) New integron-associated gene cassette encoding a 3-*N*-aminoglycoside acetyltransferase. *Antimicrob Agents Ch* **49**: 1238–1241.

Levings RS, Lightfoot D, Elbourne LD, Djordjevic SP & Hall RM (2006) New integron-associated gene cassette encoding a trimethoprim-resistant DfrB-type dihydrofolate reductase. *Antimicrob Agents Ch* **50**: 2863–2865.

Liebert CA, Hall RM & Summers AO (1999) Transposon Tn*21*, flagship of the floating genome. *Microbiol Mol Biol R* **63**: 507–522.

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

Machado E, Ferreira J, Novais A, Peixe L, Canton R, Baquero F & Coque TM (2007) Preservation of integron types among *Enterobacteriaceae* producing extended-spectrum β-lactamases in a Spanish hospital over a 15-year period (1988 to 2003). *Antimicrob Agents Ch* **51**: 2201–2204.

Markham NR & Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* **33**: W577–W581.

Markham NR & Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. *Bioinformatics Volume II. Structure, Functions and Applications* (Keith JM, ed), pp. 3–31. Humana Press, Totowa, NJ.

Márquez C, Labbate M, Ingold AJ, Roy Chowdhury P, Ramírez MS, Centrón D, Borthagaray G & Stokes HW (2008) Recovery of a functional class 2 integron from an *Escherichia coli* strain mediating a urinary tract infection. *Antimicrob Agents Ch* **52**: 4153–4154.

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

Michael GB, Cardoso M & Schwarz S (2008) Molecular analysis of multiresistant porcine *Salmonella enterica* subsp. *enterica* serovar Bredeney isolates from Southern Brazil: identification of resistance genes, integrons and a group II intron. *Int J Antimicrob Ag* **32**: 120–129.

Minakhina S, Kholodii G, Mindlin S, Yurieva O & Nikiforov V (1999) Tn*5053* family transposons are *res* site hunters sensing plasmidal *res* sites occupied by cognate resolvases. *Mol Microbiol* **33**: 1059–1068.

Naas T, Mikami Y, Imai T, Poirel L & Nordmann P (2001) Characterization of In53, a class 1 plasmid- and composite transposon-located integron of *Escherichia coli* which carries an unusual array of gene cassettes. *J Bacteriol* **183**: 235–249.

Nesvera J, Hochmannova J & Patek M (1998) An integron of class 1 is present on the plasmid pCG4 from gram-positive bacterium *Corynebacterium glutamicum*. *FEMS Microbiol Lett* **169**: 391–395.

Novick RP, Clowes RC, Cohen SN, Curtiss R, Datta N III & Falkow S (1976) Uniform nomenclature for bacterial plasmids: a proposal. *Bacteriol Rev* **40**: 168–189.

O'Mahony R, Quinn T, Drudy D, Walsh C, Whyte P, Mattar S & Fanning S (2006) Antimicrobial resistance in nontyphoidal *Salmonella* from food sources in Colombia: evidence for an unusual plasmid-localized class 1 integron in serotypes Typhimurium and Anatum. *Microb Drug Resist* **12**: 269–277.

Papagiannitsis CC, Tzouvelekis LS & Miriagou V (2008) Relative strengths of the class 1 integron promoter hybrid 2 and the combinations of strong and hybrid 1 with an active P2 promoter. *Antimicrob Agents Ch* **53**: 277–280.

Partridge SR & Hall RM (2003) The IS*1111* family members IS*4321* and IS*5075* have subterminal inverted repeats and target the terminal inverted repeats of Tn*21* family transposons. *J Bacteriol* **185**: 6371–6384.

Partridge SR & Hall RM (2005) Correctly identifying the streptothricin resistance gene cassette. *J Clin Microbiol* **43**: 4298–4300.

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

Partridge SR, Recchia GD, Stokes HW & Hall RM (2001) Family of class 1 integrons related to In4 from Tn*1696*. *Antimicrob Agents Ch* **45**: 3014–3020.

Partridge SR, Brown HJ & Hall RM (2002a) Characterization and movement of the class 1 integron known as Tn*2521* and Tn*1405*. *Antimicrob Agents Ch* **46**: 1288–1294.

Partridge SR, Collis CM & Hall RM (2002b) Class 1 integron containing a new gene cassette, *aadA10*, associated with Tn*1404* from R151. *Antimicrob Agents Ch* **46**: 2400–2408.

Patzer J, Toleman MA, Deshpande LM, Kaminska W, Dzierzanowska D, Bennett PM, Jones RN & Walsh TR (2004) *Pseudomonas aeruginosa* strains harbouring an unusual $bla_{\text{VIM-4}}$ gene cassette isolated from hospitalized children in Poland (1998–2001). *J Antimicrob Chemoth* **53**: 451–456.

Peters JE & Craig NL (2001) Tn*7*: smarter than we thought. *Nat Rev Mol Cell Bio* **2**: 806–814.

Ploy MC, Denis F, Courvalin P & Lambert T (2000) Molecular characterization of integrons in *Acinetobacter baumannii*: description of a hybrid class 2 integron. *Antimicrob Agents Ch* **44**: 2684–2688.

Poirel L, Le Thomas I, Naas T, Karim A & Nordmann P (2000) Biochemical sequence analyses of GES-1, a novel class A extended-spectrum β-lactamase, and the class 1 integron In52 from *Klebsiella pneumoniae*. *Antimicrob Agents Ch* **44**: 622–632.

Poirel L, Lambert T, Turkoglu S, Ronco E, Gaillard J & Nordmann P (2001) Characterization of class 1 integrons from *Pseudomonas aeruginosa* that contain the $bla_{\text{VIM-2}}$ carbapenem-hydrolyzing β-lactamase gene and of two novel aminoglycoside resistance gene cassettes. *Antimicrob Agents Ch* **45**: 546–552.

Poirel L, Briñas L, Fortineau N & Nordmann P (2005) Integron-encoded GES-type extended-spectrum β-lactamase with

increased activity toward aztreonam in *Pseudomonas aeruginosa*. *Antimicrob Agents Ch* **49**: 3593–3597.

Post V & Hall RM (2008) Insertion sequences in the IS*1111* family that target the *attC* recombination sites of integron-associated gene cassettes. *FEMS Microbiol Lett* **290**: 182–187.

Post V, Recchia GD & Hall RM (2007) Detection of gene cassettes in Tn*402*-like class 1 integrons. *Antimicrob Agents Ch* **51**: 3467–3468.

Pournaras S, Ikonomidis A, Tzouvelekis LS, Tokatlidou D, Spanakis N, Maniatis AN, Legakis NJ & Tsakris A (2005) VIM-12, a novel plasmid-mediated metallo-β-lactamase from *Klebsiella pneumoniae* that resembles a VIM-1/VIM-2 hybrid. *Antimicrob Agents Ch* **49**: 5153–5156.

Quiroga C, Roy PH & Centron D (2008) The *Sma*I2 class C group II intron inserts at integron *attC* sites. *Microbiology* **154**: 1341–1353.

Radstrom P, Sköld O, Swedberg G, Flensburg J, Roy PH & Sundström L (1994) Transposon Tn*5090* of plasmid R751, which carries an integron, is related to Tn*7*, Mu, and the retroelements. *J Bacteriol* **176**: 3257–3268.

Ramirez MS, Quiroga C & Centron D (2005) Novel rearrangement of a class 2 integron in two non-epidemiologically related isolates of *Acinetobacter baumannii*. *Antimicrob Agents Ch* **49**: 5179–5181.

Ramirez MS, Parenteau TR, Centron D & Tolmasky ME (2008) Functional characterization of Tn*1331* gene cassettes. *J Antimicrob Chemoth* **62**: 669–673.

Rather PN, Munayyer H, Mann PA, Hare RS, Miller GH & Shaw KJ (1992) Genetic analysis of bacterial acetyltransferases: identification of amino acids determining the specificities of the aminoglycoside 6′-*N*-acetyltransferase Ib and IIa proteins. *J Bacteriol* **174**: 3196–3203.

Recchia GD & Hall RM (1995a) Plasmid evolution by acquisition of mobile gene cassettes: plasmid pIE723 contains the *aadB* gene cassette precisely inserted at a secondary site in the IncQ plasmid RSF1010. *Mol Microbiol* **15**: 179–187.

Recchia GD & Hall RM (1995b) Gene cassettes: a new class of mobile element. *Microbiology* **141**: 3015–3027.

Recchia GD & Hall RM (1997) Origins of the mobile gene cassettes found in integrons. *Trends Microbiol* **5**: 389–394.

Recchia GD & Sherratt DJ (2002) Gene acquisition in bacteria by integron-mediated site-specific recombination. *Mobile DNA II* (Craig NL, Craigie R, Gellert M & Lambowitz AM, eds), pp. 162–176. ASM Press, Washington, DC.

Robicsek A, Strahilevitz J, Jacoby GA, Macielag M, Abbanat D, Park CH, Bush K & Hooper DC (2006) Fluoroquinolone-modifying enzyme: a new adaptation of a common aminoglycoside acetyltransferase. *Nat Med* **12**: 83–88.

Rowe-Magnus DA & Mazel D (2002) The role of integrons in antibiotic resistance gene capture. *Int J Med Microbiol* **292**: 115–125.

Rowe-Magnus DA, Guerout AM, Ploncard P, Dychinco B, Davies J & Mazel D (2001) The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *P Natl Acad Sci USA* **98**: 652–657.

Rowe-Magnus DA, Guerout AM & Mazel D (2002) Bacterial resistance evolution by recruitment of super-integron gene cassettes. *Mol Microbiol* **43**: 1657–1669.

Scoulica EV, Neonakis IK, Gikas AI & Tselentis YJ (2004) Spread of $bla_{\text{VIM-1}}$-producing *E. coli* in a university hospital in Greece. Genetic analysis of the integron carrying the $bla_{\text{VIM-1}}$ metallo-β-lactamase gene. *Diagn Micr Infec Dis* **48**: 167–172.

Searls DB (2002) The language of genes. *Nature* **420**: 211–217.

Senda K, Arakawa Y, Ichiyama S, Nakashima K, Ito H, Ohsuka S, Shimokata K, Kato N & Ohta M (1996) PCR detection of metallo-β-lactamase gene (bla_{IMP}) in gram-negative rods resistant to broad-spectrum β-lactams. *J Clin Microbiol* **34**: 2909–2913.

Shaw KJ, Rather PN, Hare RS & Miller GH (1993) Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiol Rev* **57**: 138–163.

Shi L, Zheng M, Xiao Z, Asakura M, Su J, Li L & Yamasaki S (2006) Unnoticed spread of class 1 integrons in gram-positive clinical strains isolated in Guangzhou, China. *Microbiol Immunol* **50**: 463–467.

Smith CA, Caccamo M, Kantardjieff KA & Vakulenko S (2007) Structure of GES-1 at atomic resolution: insights into the evolution of carbapenamase activity in the class A extended-spectrum β-lactamases. *Acta Crystallogr D* **63**: 982–992.

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

Stokes HW & Hall RM (1992) The integron In1 in plasmid R46 includes two copies of the *oxa2* gene cassette. *Plasmid* **28**: 225–234.

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

Stokes HW, Nesbo CL, Holley M, Bahl MI, Gillings MR & Boucher Y (2006) Class 1 integrons potentially predating the association with Tn*402*-like transposition genes are present in a sediment microbial community. *J Bacteriol* **188**: 5722–5730.

Sunde M (2005) Class I integron with a group II intron detected in an *Escherichia coli* strain from a free-range reindeer. *Antimicrob Agents Ch* **49**: 2512–2514.

Szekeres S, Dauti M, Wilde C, Mazel D & Rowe-Magnus DA (2007) Chromosomal toxin-antitoxin loci can diminish large-scale genome reductions in the absence of selection. *Mol Microbiol* **63**: 1588–1605.

Tetu SG & Holmes AJ (2008) A family of insertion sequences that impacts integrons by specific targeting of gene cassette recombination sites, the IS*1111-attC* group. *J Bacteriol* **190**: 4959–4970.

Toleman MA, Bennett PM & Walsh TR (2006) IS*CR* elements: novel gene-capturing systems of the 21st century? *Microbiol Mol Biol R* **70**: 296–316.

Toro N, Jimenez-Zurdo JI & Garcia-Rodriguez FM (2007) Bacterial group II introns: not just splicing. *FEMS Microbiol Rev* **31**: 342–358.

Tran van Nhieu G & Collatz E (1987) Primary structure of an aminoglycoside 6′-*N*-acetyltransferase AAC(6′)-4, fused *in vivo* with the signal peptide of the Tn*3*-encoded β-lactamase. *J Bacteriol* **169**: 5708–5714.

Vanhoof R, Hannecart-Pokorni E & Content J (1998) Nomenclature of genes encoding aminoglycoside-modifying enzymes. *Antimicrob Agents Ch* **42**: 483.

Walsh TR (2008) Clinically significant carbapenemases: an update. *Curr Opin Infect Dis* **21**: 367–371.

Walther-Rasmussen J & Hoiby N (2007) Class A carbapenemases. *J Antimicrob Chemoth* **60**: 470–482.

Weldhagen GF, Kim B, Cho C-H & Lee SH (2006) Definitive nomencalture of GES/IBC-type extended-spectrum β-lactamases. *J Microbiol Biotech* **16**: 1837–1840.

White PA, McIver CJ, Deng Y & Rawlinson WD (2000) Characterisation of two new gene cassettes, *aadA5* and *dfrA17*. *FEMS Microbiol Lett* **182**: 265–269.

White PA, McIver CJ & Rawlinson WD (2001) Integrons and gene cassettes in the *Enterobacteriaceae*. *Antimicrob Agents Ch* **45**: 2658–2661.

Xu H, Davies J & Miao V (2007) Molecular characterization of class 3 integrons from *Delftia* spp. *J Bacteriol* **189**: 6276–6283.

Yano H, Kuga A, Okamoto R, Kitasato H, Kobayashi T & Inoue M (2001) Plasmid-encoded metallo-β-lactamase (IMP-6) conferring resistance to carbapenems, especially meropenem. *Antimicrob Agents Ch* **45**: 1343–1348.

Yatsuyanagi J, Saito S, Konno T, Harata S, Suzuki N & Amano K (2005) The ORF1 gene located on the class-1-integron-associated gene cassette actually represents a novel fosfomycin resistance determinant. *Antimicrob Agents Ch* **49**: 2573.

Yu HS, Lee JC, Kang HY *et al.* (2003) Changes in gene cassettes of class 1 integrons among *Escherichia coli* isolates from urine specimens collected in Korea during the last two decades. *J Clin Microbiol* **41**: 5429–5433.

Zong Z, Partridge SR & Iredell JR (2009) A bla_{VEB} variant, $bla_{VEB\text{-}6}$, associated with repeated elements in a complex genetic structurein. *Antimicrob Agents Ch* **53**: 1693–1697.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. OXA-10 variants.
Table S2. OXA-13 variants.
Table S3. OXA-2 variants.
Table S4. VEB variants.
Table S5. IMP-1 variants.
Table S6. IMP-2 variants.
Table S7. VIM-1 variants.
Table S8. VIM-2 variants.
Table S9. *aadA1*/*aadA2* hybrid cassettes.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.