Gene cassettes: a new class of mobile element

Gavin D. Recchia¹² and Ruth M. Hall¹

Author for correspondence: Ruth M. Hall. Tel: +61 2 886 4962. Fax: +61 2 886 4805.
e-mail: hall@pelican.dbe.csiro.au

¹CSIRO Division of Biomolecular Engineering, Sydney Laboratory, PO Box 184, North Ryde, NSW 2113, Australia
²School of Biological Sciences, Macquarie University, NSW 2109, Australia

Keywords: gene cassette, integron, site-specific recombination, antibiotic-resistance genes

Overview

Gene cassettes form a diverse group of small mobile elements that already includes over 40 members. Usually, each cassette contains only a single gene and a specific recombination site. At present, the genes found in cassettes are mostly antibiotic-resistance genes, but it is likely that gene cassettes represent a more general modular gene-packaging mechanism. These gene cassettes are normally found integrated at a specific site in an integron, where they are often part of multi-cassette arrays. Three classes of integron are currently known, all of which appear to be able to integrate the same set of cassettes. However, cassettes are also found rarely at non-specific sites. They move by conservative site-specific recombination catalysed by an integrase encoded by the recipient integron, and thus form a family of mobile elements that are more akin to integrating phage and plasmids than to transposons. Cassettes do not, in general, include a promoter and are transcribed from a promoter in the integron. The multi-component cassette–integron system is thus a natural cloning and expression system.

Historical perspective

Mobile gene cassettes were first identified as the various integrated antibiotic-resistance genes found in integrons. In the early 1980s, studies using restriction mapping and heteroduplex analysis revealed that different sets of antibiotic-resistance genes were sometimes found in the same location in otherwise closely related plasmids such as R388 and pSa (Ward & Grinstead, 1982) or transposons such as Tn21, Tn2003 and Tn2424 (Schmidt & Klopf-Kaul, 1984; Tanaka et al., 1983; Meyer et al., 1983). As the sequences of antibiotic-resistance genes began to be determined in the mid-1980s, several of them including dfrB2 (dhfrII) from R388, aadA2 from pSa, aadA1 from Tn21, oxyA from Tn2003 and oxyA2 from R46 were found to be flanked by identical sequences, indicating that they were all integrated at the same site within a common structure (Cameron et al., 1986; Hall & Vockler, 1987; Ouellette et al., 1987). Over the ensuing years, many further genes determining resistance to aminoglycosides (gentamicin, tobramycin, amikacin, kanamycin, streptomycin and spectinomycin), trimethoprim, chloramphenicol and β-lactam antibiotics were also found in the same site and a second site in the transposon Tn7 was identified (see Hall et al., 1991). There are several cassettes encoding distinct determinants of resistance to each of these antibiotics or antibiotic families. Further cassettes include open reading frames (ORFs), whose functions have not yet been identified (see Hall et al., 1991). At the present time, over 40 distinct cassettes have been identified.

These integrated cassettes are often part of an array that contains two or more cassettes, and as sufficient sequences became available it was possible to determine accurately the boundaries of the gene cassettes (Hall et al., 1991) and to deduce that they were discrete genetic units, integrated at a unique site in larger mobile units that were named integrons (Stokes & Hall, 1989). A gene encoding a product related to integrases of the 7 integrase family of site-specific recombinases was identified in the conserved sequences located 5' to the cassettes (5'-CS) in members of the major integron family (Hall & Vockler, 1987; Ouellette & Roy, 1987), and this integrase was shown to catalyse site-specific recombination (Martinez & de la Cruz, 1988, 1990). Subsequent studies demonstrated that the gene cassettes were indeed mobile elements that can be excised from or integrated into an integron by this integrase (Collis & Hall, 1992a, b; Collis et al., 1993; Bunny et al., 1995).

Diversity of cassette-encoded genes

The gene cassettes for which sequences are available are listed in Table 1. Cassettes are generally identified by the name of the gene they encode, or in the case of cassettes containing ORFs whose function is unknown, the ORFs have been assigned letters in the order of their identi-
<table>
<thead>
<tr>
<th>Gene cassette*</th>
<th>Protein</th>
<th>GenBank/EMBL accession no.†</th>
<th>Position of cassette‡,§</th>
<th>Length of cassette (bp)§</th>
<th>Length of coding region (bp)¶</th>
<th>Length of 59-base element (bp)§</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resistance to β-lactams</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Class A β-lactamases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>bla</em>P1</td>
<td>PSE-1/CARB-2**</td>
<td>Z18955</td>
<td>102–1145</td>
<td>1044</td>
<td>915</td>
<td>111</td>
</tr>
<tr>
<td><em>bla</em>P2</td>
<td>–</td>
<td>D13210</td>
<td>163–1206</td>
<td>1044</td>
<td>915</td>
<td>111</td>
</tr>
<tr>
<td><em>bla</em>P3</td>
<td>CARB-4</td>
<td>U14749</td>
<td>129–1151</td>
<td>&gt;1023</td>
<td>867</td>
<td>&gt;92</td>
</tr>
<tr>
<td><strong>Class B β-lactamases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>bla</em>IMP</td>
<td>IMP-1</td>
<td>D50438</td>
<td>1179–2058</td>
<td>880</td>
<td>741</td>
<td>127</td>
</tr>
<tr>
<td><strong>Class D β-lactamases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>oxa</em>1</td>
<td>OXA-1</td>
<td>J02967</td>
<td>1289–2292</td>
<td>1004</td>
<td>831</td>
<td>90</td>
</tr>
<tr>
<td><em>oxa</em>2</td>
<td>OXA-2</td>
<td>M95287</td>
<td>1439–2314</td>
<td>876</td>
<td>828</td>
<td>70</td>
</tr>
<tr>
<td><em>oxa</em>3</td>
<td>OXA-3</td>
<td>L07945</td>
<td>40–900</td>
<td>861</td>
<td>828</td>
<td>&gt;56</td>
</tr>
<tr>
<td><em>oxa</em>5</td>
<td>OXA-5</td>
<td>X58272</td>
<td>63–977</td>
<td>915</td>
<td>804</td>
<td>106</td>
</tr>
<tr>
<td><em>oxa</em>7</td>
<td>OXA-7</td>
<td>X75562</td>
<td>115–988</td>
<td>874</td>
<td>801</td>
<td>65</td>
</tr>
<tr>
<td><em>oxa</em>9</td>
<td>OXA-9</td>
<td>M55547</td>
<td>2269–3225</td>
<td>957</td>
<td>840</td>
<td>69</td>
</tr>
<tr>
<td><em>oxa</em>10</td>
<td>OXA-10 (PSE-2)</td>
<td>U37105</td>
<td>102–1021</td>
<td>920</td>
<td>801</td>
<td>111</td>
</tr>
<tr>
<td><strong>Resistance to aminoglycosides</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aminoglycoside adenylyltransferases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>aad</em>A1a</td>
<td>AAD(3*)</td>
<td>X12870</td>
<td>1290–2145</td>
<td>856</td>
<td>792</td>
<td>60</td>
</tr>
<tr>
<td><em>aad</em>A1b</td>
<td>AAD(3*)</td>
<td>M95287</td>
<td>2315–3170</td>
<td>856</td>
<td>792</td>
<td>60</td>
</tr>
<tr>
<td><em>aad</em>A2</td>
<td>AAD(3*)</td>
<td>X68227</td>
<td>145–1000</td>
<td>856</td>
<td>780</td>
<td>60</td>
</tr>
<tr>
<td><em>aad</em>B</td>
<td>AAD(2*)</td>
<td>L06418</td>
<td>1287–1877</td>
<td>591</td>
<td>534</td>
<td>60</td>
</tr>
<tr>
<td>Aminoglycoside acetyltransferases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>aac</em>A1††</td>
<td>AAC(6')-Ia</td>
<td>M18967</td>
<td>713–1487</td>
<td>&gt;778</td>
<td>558</td>
<td>&gt;††</td>
</tr>
<tr>
<td><em>aac</em>A4</td>
<td>AAC(6')-Ib</td>
<td>M55547</td>
<td>820–1456</td>
<td>637</td>
<td>555</td>
<td>70</td>
</tr>
<tr>
<td><em>aac</em>A (orfB)</td>
<td>AAC(6')-Id</td>
<td>X12618</td>
<td>896–1421</td>
<td>526</td>
<td>450</td>
<td>72</td>
</tr>
<tr>
<td><em>aac</em>A7</td>
<td>AAC(6')-II</td>
<td>U13880</td>
<td>299–889</td>
<td>591</td>
<td>459</td>
<td>112</td>
</tr>
<tr>
<td><em>aac</em>A</td>
<td>AAC(6')-Ia</td>
<td>M29695</td>
<td>682–1309</td>
<td>628</td>
<td>555</td>
<td>60</td>
</tr>
<tr>
<td><em>aac</em>A</td>
<td>AAC(6')-Ib</td>
<td>L06163</td>
<td>507–1159</td>
<td>653</td>
<td>543</td>
<td>97</td>
</tr>
<tr>
<td><em>aac</em>C1</td>
<td>AAC(3')-Ia</td>
<td>X13852</td>
<td>1304–1880</td>
<td>577</td>
<td>465</td>
<td>109</td>
</tr>
<tr>
<td><em>aac</em>C</td>
<td>AAC(3')-Ib</td>
<td>L06157</td>
<td>609–1106</td>
<td>&gt;498</td>
<td>465</td>
<td>&gt;34</td>
</tr>
<tr>
<td><strong>Resistance to chloramphenicol</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chloramphenicol acetyltransferases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cat</em>B2</td>
<td>CATB2</td>
<td>M80188</td>
<td>298–1036</td>
<td>739</td>
<td>633</td>
<td>72</td>
</tr>
<tr>
<td><em>cat</em>B3</td>
<td>CATB3</td>
<td>U13880</td>
<td>890–1604</td>
<td>715</td>
<td>633</td>
<td>60</td>
</tr>
<tr>
<td><em>cat</em>B5</td>
<td>CATB5</td>
<td>X82455</td>
<td>2–678</td>
<td>&gt;677</td>
<td>633</td>
<td>&gt;25</td>
</tr>
<tr>
<td>Chloramphenicol exporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cml</em>A</td>
<td>CmlA</td>
<td>U12338</td>
<td>3227–4775</td>
<td>1549</td>
<td>1260</td>
<td>70</td>
</tr>
<tr>
<td><strong>Resistance to trimethoprim</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class A dihydrofolate reductases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>dfr</em>A1</td>
<td>DHFRIa</td>
<td>X09026</td>
<td>216–792</td>
<td>577</td>
<td>474</td>
<td>95</td>
</tr>
<tr>
<td><em>dfr</em>A5</td>
<td>DHFRV</td>
<td>X12868</td>
<td>1287–1854</td>
<td>568</td>
<td>474</td>
<td>87</td>
</tr>
<tr>
<td><em>dfr</em>A7</td>
<td>DHFRVII</td>
<td>X58425</td>
<td>573–1189</td>
<td>617</td>
<td>474</td>
<td>134</td>
</tr>
<tr>
<td><em>dfr</em>A12</td>
<td>DHFRXII</td>
<td>Z21672</td>
<td>302–885</td>
<td>584</td>
<td>498</td>
<td>90</td>
</tr>
<tr>
<td><em>dfr</em>A14</td>
<td>DHFR Ib</td>
<td>S76821</td>
<td>216–738</td>
<td>&gt;523</td>
<td>483</td>
<td>&gt;43</td>
</tr>
<tr>
<td>Class B dihydrofolate reductases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>dfr</em>B1</td>
<td>DHFRIa</td>
<td>U36276</td>
<td>573–1057</td>
<td>485</td>
<td>237</td>
<td>57</td>
</tr>
<tr>
<td><em>dfr</em>B2</td>
<td>DHFRIb</td>
<td>J01773</td>
<td>707–1090</td>
<td>384</td>
<td>237</td>
<td>57</td>
</tr>
<tr>
<td><em>dfr</em>B3</td>
<td>DHFRIc</td>
<td>X72585</td>
<td>5854–6261</td>
<td>408</td>
<td>237</td>
<td>57</td>
</tr>
<tr>
<td><strong>Resistance to streptothricin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptothricin acetyltransferase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>sat</em></td>
<td>SAT-2</td>
<td>X15995</td>
<td>247–830</td>
<td>584</td>
<td>525</td>
<td>60</td>
</tr>
</tbody>
</table>
Table 1 (cont.)

<table>
<thead>
<tr>
<th>Gene cassette*</th>
<th>Protein</th>
<th>GenBank/EMBL accession no.†</th>
<th>Position of cassette‡,§</th>
<th>Length of cassette (bp)¶</th>
<th>Length of coding region (bp)§</th>
<th>Length of 59-base element (bp)¶</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resistance to antiseptics and disinfectants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quaternary ammonium compound exporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qacE</td>
<td>QacE</td>
<td>X72585</td>
<td>4947–5533</td>
<td>587</td>
<td>333</td>
<td>141</td>
</tr>
<tr>
<td><strong>Unidentified ORFs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>orfA</td>
<td>J01773</td>
<td>1091–1167</td>
<td>501</td>
<td>435</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>orfC</td>
<td>X12869</td>
<td>7–430</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>orfD</td>
<td>M95287</td>
<td>1150–1656</td>
<td>507</td>
<td>378</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>orfE</td>
<td>U12338</td>
<td>2109–2370</td>
<td>262</td>
<td>246</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>orfF</td>
<td>Z21672</td>
<td>886–1205</td>
<td>320</td>
<td>291</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

* Cassettes are named after the gene they encode. However, the gene nomenclature used here may differ from that found in the original publications, and in many cases the genes have not previously been assigned names. As an agreed numbering system for the aacA and aacC genes is not currently available, several of these genes are not numbered. The dfrA and dfrB genes encode proteins belonging to two distinct families. The ORFs with no known function were assigned letters in the order of their publication.

† The first or the first accurate complete cassette sequence has been used in each case. Incomplete sequences have been used only where no complete cassette sequence is available.

‡ Positions of cassettes are frequently not or incorrectly identified in the published sequence and database entries.

§ > Database entry ends before the end of the cassette.

¶ The first in-frame initiation codon within the cassette boundary has been assumed to be the start codon.

** The genes encoding the PSE-4 and CARB-3 β-lactamases each differ by only one nucleotide from the blap1 sequence.

†† The aacA1 gene is part of a cassette that also contains an ORF, orfG, followed by a 59-base element (A. Gravel, R. Parent & P. H. Roy, personal communication).

‡‡ The complete sequence of the orfA cassette is not present in any single database entry and has therefore been compiled from two separate entries.

§§ Coding region extends beyond the cassette boundary.

Mobile gene cassettes

fication and the cassettes are identified by these names (see Hall et al., 1991). Table 1 lists the boundaries of the cassettes, many of which were not identified or were incorrectly identified in the primary publication of the sequence. The extent of each gene, assuming that the first in-frame initiation codon within the cassette boundary is the start codon, is also listed in Table 1. The genes listed account for a significant proportion of the antibiotic-resistance genes found in Gram-negative bacteria of the Enterobacteriaceae family and the genus Pseudomonas. For most antibiotics, several quite distinct resistance genes are found, and in some cases the mechanism by which resistance is conferred is different. For example, resistance to chloramphenicol by acetylation (aac genes) and efflux of the antibiotic from the cells (multidrug resistance efflux systems) are both represented, and for the aminoglycosides, modification of the antibiotic by acetylation (aacA and aacC genes) and adenylylation (aadA and aadB genes) are both represented. The β-lactamases encoded by cassettes fall into three distinct families, class A (blaP genes), class B metallo-β-lactamases (bld genes), class D (oxa genes), and the dihydrofolate reductases conferring trimethoprim resistance are also from two unrelated protein families (dfrA and dfrB genes). Partial sequences of several further potential cassettes (not listed) are found flanking some known cassette sequences, and further cassettes are likely to be identified in the future. Though at present the genes found in cassettes are mostly antibiotic-resistance genes, this bias is likely to reflect the origins of the genes studied, which are from antibiotic-resistant clinical isolates.

Structure of gene cassettes

Cassettes vary considerably in total length from 262 to 1549 bp (see Fig. 1, Table 1). This is largely because of differences in the size of the gene, as in most cassettes little non-coding sequence flanks the gene. Despite the diversity in the size and function of the genes they contain, cassettes share common features (Hall et al., 1991). Each cassette includes a single gene that covers most of the cassette length and a recombination site known as a 59-base element located downstream of the gene (Fig. 1). Though the 59-base elements differ substantially in length from 57 to 141 bp (see Table 1) they share common features (see below). In many cases, a 59-base element was not identified in the primary publications of the sequences of the cassettes, but in all cases where the complete cassette sequence is available a 59-base element has been subsequently identified (Hall et al., 1991; G. D. Recchia & R. M. Hall, unpublished observations) and their precise locations are listed in Table 1.
Resistance to β-lactams

Class A β-lactamases

\( \text{blaP1} \)

\( \text{blaP2} \)

\( \text{blaP3} \)

Class B β-lactamases

\( \text{blaIMP} \)

Class C β-lactamases

\( \text{oxa1} \)

\( \text{oxa2} \)

\( \text{oxa3} \)

\( \text{oxa5} \)

Resistance to trimethoprim

Dihydrofolate reductases

\( \text{dfrA1} \)

\( \text{dfrA5} \)

\( \text{dfrA7} \)

\( \text{dfrA12} \)

\( \text{dfrA14} \)

Resistance to chloramphenicol

Chloramphenicol acetyltransferases

\( \text{catB2} \)

\( \text{catB3} \)

\( \text{catB5} \)

Chloramphenicol exporter

\( \text{cmiA} \)

Fig. 1. For legend see facing page.
Mobile gene cassettes

Resistance to aminoglycosides
Aminoglycoside adenylyltransferases

- **aadA1a**
- **aadA1b**
- **aadA2**
- **aadB**

Aminoglycoside acetyltransferases

- **aacA1**
- **aacA4**
- **aacA (lla)**
- **aacA (llb)**
- **aac7**
- **aacA (orfB)**
- **aacCl**
- **aacC (llb)**

Resistance to antiseptics and disinfectants
Quaternary ammonium compound exporter

- **qacE**
- **sat**

Resistance to streptothricin
Streptothricin acetyltransferase

- **Nd**

Unidentified ORFs

- **orfA**
- **orfC**
- **orfD**
- **orfE**
- **orfF**

While gene cassettes are most commonly found in an integrated form either in an integron or at a non-specific location, they can also exist as covalently closed circular molecules (Collis & Hall, 1992b). Free circular cassettes are formed by excision of the cassette from an integron, and are likely to be important intermediates in the dissemination of cassettes (see Fig. 2).

Because of the nature of the site-specific recombination event that leads to integration of a free circular cassette, a small part of the cassette-associated 59-base element is found at the start of the linearized cassette and the remainder is downstream (at the 3' end) of the gene (Fig. 3). The conserved triplet GTT, which is part of the core site GTTRRRY (R = purine, Y = pyrimidine) found at the 3' end of 59-base elements, can be found at both ends of integrated cassette sequences, though only one GTT is part of the cassette (Hall et al., 1991). This led initially to some ambiguity in defining the precise boundaries of cassettes, which has now been resolved experimentally by determining the precise location of the crossover in recombination events involving one partner with a mutation in the GTT (D. O'Gorman, H. W. Stokes & R. M. Hall, unpublished observations). The crossover occurs between the G and the first T, and thus the first six bases of each integrated cassette are derived from the 59-base element and have the sequence TTRRRY. The last base of the integrated cassette sequence is conserved and is a G residue (see Fig. 3).

The cassette-associated recombination sites

An essential feature of gene cassettes is the presence of a recombination site known as a 59-base element that is located downstream of the gene coding region (Hall et al., 1991). Generally, each gene cassette possesses a unique 59-base element, and both the length and sequence of 59-base elements can vary considerably (see Table 1 and Fig. 4). However, despite this heterogeneity, all 59-base elements share several features. They are bounded by an inverse core site (RYYYAAC) at the end closest to the 3' end of the gene coding region and a core site (GTTRRRY) at the other end, and are related to a consensus sequence that covers approximately the first and last 20 bases (Collis & Hall, 1992b). Most 59-base elements have a central axis of symmetry as the consensus regions are imperfect inverted repeats. All 59-base elements also contain inverted repeat sequences in their central region, and it is this region that varies in length.

Within the 59-base element family, a few groups of closely related elements can be identified. Members of each group are of similar length and share considerable but not complete sequence identity with one another. The largest group of 59-base elements includes those first identified by Cameron et al. (1986) and now contains 12 members, several of which are associated with completely unrelated genes, e.g. the adaA genes, adaB, catB3 and orfD. Most of the 59-base elements in this group differ from the 60 base consensus for the group (Hall et al., 1991) by only a few bases. However, over half of the known 59-base elements, apart from conforming to the consensus sequences for all 59-base elements at their outer ends, are not otherwise closely related to any other known 59-base element. Nonetheless, several 59-base elements with quite different sequences and lengths have been shown to function as recombination sites recognized by the integron integrase (IntI) (Martinez & de la Cruz, 1990; Hall et al., 1991;
Fig. 4. Sequences of the 59-base elements of the aadB, aacA4 and dfrA7 cassettes. The sequences shown are those present in the circular form of each cassette and are shown folded only in order to emphasize the inverted repeat structure. The aadB, aacA4 and dfrA7 59-base elements are 60, 70 and 134 bases long, respectively, yet each retains the features common to all 59-base elements. The inverse core site at the 5’ end and the core site at the 3’ end of each 59-base element are shown in bold type.

Collis & Hall, 1992a; Bunny et al., 1995; H. W. Stokes & R. M. Hall, unpublished observations). Thus, despite their diversity, they are indeed members of a family of functional recombination sites that are all recognized by the same integrase(s).

Fused cassettes

Though gene cassettes normally contain a single gene coding region and a 59-base element, an examination of published sequences reveals several instances in which two potential coding regions are found within a single cassette and it seems likely that at least some of these cassettes were created by the fusion of two cassettes each containing a single gene. Two potential routes which could give rise to fused cassettes can be envisaged (see Fig. 5). Firstly, a cassette fusion would be generated by any deletion event with end-points in two adjacent gene cassettes. The catB4A7-aacA4 cassette fusion (see Bunny et al., 1995 for details) is likely to have arisen by this route. The 3’ end of the catB4 gene, the catB4 59-base element and the 5’ non-coding region of the aacA4 gene cassette have been deleted to produce a single cassette containing the truncated catB4A7 gene and the complete aacA4 gene.

Similarly, the relationships between the orfD, orfE and orfF cassettes may be explained by a cassette fusion (Bunny et al., 1995).

Fused cassettes that retain both complete genes may be created by any event involving loss of part or all of the 59-base element from one cassette leading to its fusion with a second downstream cassette. This appears to have occurred in the case of the aadA1-oxa9 cassette fusion (Tolmasky, 1990). The 3’ end of the aadA1 gene coding region is separated from the start of the oxa9 cassette by only 14 bp and the 59-base element normally associated with the aadA1 gene is not present. The only complete 59-base element is found downstream of the oxa9 gene. A potentially similar case is that of the aacA1 gene which does not appear to have a 59-base element associated with it, and is followed 4 bp downstream by a further ORF (Tenover et al., 1988). This ORF, orfG, is 142 codons long and is followed by a 59-base element (A. Gravel, R. Parent & P. H. Roy, personal communication). However, to date neither aacA1 nor orfG have been found as single gene cassettes. Loss of a 59-base element could occur in a variety of ways, e.g. slippage during DNA replication caused by a stem–loop structure in the single-stranded 59-base element DNA template, and fusions of this type may play a role in creating permanent new gene arrays, or operons.

Fig. 5. Cassette fusions: two potential routes for fusion of adjacent cassettes. The genes (numbered 1 and 2) are represented by open boxes and their 59-base elements are represented by filled boxes. (a) A deletion event with end-points in each cassette resulting in the truncation of one or both genes. (b) Loss of the 59-base element from one cassette resulting in the retention of both gene coding regions.
Locations of integrated cassettes

Most of the cassettes listed in Table 1 were found integrated at a specific site in an integron of the type shown in Fig. 6(a). Integrons of this class (class 1) are mobile elements that include the gene for the IntI integrase and the cassette integration site attI (see below for details). Most of the class 1 integrons whose structures have been examined in detail are defective transposons or transposon remnants (Hall et al., 1994; H. J. Brown, H. W. Stokes & R. M. Hall, unpublished observations), though Tn402 is an integron that is an active transposon (Shapiro & Sporn, 1977; Radstrom et al., 1994). The cassettes are located at a unique position within the integron-associated recombination site, and several cassettes can be present simultaneously. Integrated cassettes are always in the same orientation, with the 5' end of the gene closest to the integron segment that contains the intI gene (see Fig. 6a), and this segment is designated the 5'-conserved segment (5'-CS). Class 1 integrons are the most prevalent class of integron found in clinical isolates and most contain a sulphonamide-resistance gene, sul1, in a conserved region found 3' to the integrated cassettes and known as the 3'-conserved segment (3'-CS).

Cassettes have also been found at a unique site near the left end of the transposon Tn7 and its close relatives (see Hall et al., 1991; Sundstrom et al., 1991; Tietze & Brevet, 1991; Young et al., 1994). The sequences flanking these cassette arrays are not related to those flanking the cassette arrays described above. However, the general organization of the left end of Tn7 (Fig. 6b) is similar to that of the class 1 integrons described above. A putative but defective intI gene (intI2 is interrupted by a termination codon), whose product is 40% identical to that of IntI1, is located 5' to the first cassette, at the left end of Tn7 (Hall & Vockler, 1987; A. Pelletier & P. H. Roy, personal communication) and Tn7 and its relatives therefore form a second class of integron.

Recently, a third class of integron has been identified (Arakawa et al., 1995). In this case, the available sequence data have revealed a new intI gene (intI3) located 5' to the blaIMP cassette which was first found in a class 1 integron (Osano et al., 1994). The putative integrase (IntI3) is 61% identical to the IntI1 integrase (Arakawa et al., 1995). Whether this integrase is also located within a transposon or defective transposon remnant, as are the class 1 and 2 integrases, remains to be established.

The cassette array found in Tn1331 (Tolmasky, 1990) includes the aacA4 and aadA1–oxa9 cassettes flanked by approximately 100 bp of the 3'-CS of class 1 integrons and a few bases that may derive from the 5'-CS. This fragment of an integron has become incorporated into a Tn3-like transposon to give rise to Tn1331.

Rarely, known cassettes have also been found outside an integron context and these cassettes (aadB and dfrA14) are

---

Fig. 6. Locations of gene cassettes. Cassettes are represented by solid lines and their associated 59-base elements are represented by filled circles. (a) Some naturally occurring class 1 integrons. The intI1 gene encoding the DNA integrase responsible for cassette movement is contained in the 5'-CS, and the 3'-CS found downstream of the integrated cassettes includes the sulphonamide-resistance gene, sul1. The attI recombination site is represented by a striped box. The origins of these integrons are: In0, pVS1; In1, R46; In2, Tn21 (NR1); In3, R388; In4, Tn1696; In6, pSa; In7, pDGO100; In30, pBVH301. (b) The left end of Tn7 and its close relatives. The putative intI2* gene, located 5' to the integrated cassettes, is interrupted by a termination codon in Tn7, Tn1825 and Tn1826 (Hall & Vockler, 1987; A. Pelletier & P. H. Roy, personal communication) and in Tn4132 (S. J. Rosser & H.-K. Young, personal communication). Transposition genes present downstream of the cassettes are also indicated.
located at non-specific sites (Recchia & Hall, 1995; H.-K. Young, personal communication). It appears that this is the result of recombination catalysed by IntI but involving a secondary recombination site (Recchia & Hall, 1995), and it is likely that further ‘wild’ cassettes will be found in the future. Indeed, one further potential example of a gene cassette located at a non-specific location is the dfrA6 gene (Wylie & Koornhof, 1991). Though the sequences flanking the dfrA6 gene do not resemble any known integron or gene cassette sequences, the DHFPRV1 protein belongs to the same subfamily as DHFRI, -V and -VII, all of which are encoded by cassette-associated genes, and a potential 59-base element containing inverted repeats and showing some similarity to the consensus for the outer ends of 59-base elements can be identified 3’ to the dfrA6 gene (see Sundström et al., 1993). However, it remains to be proven that dfrA6 is in fact part of a cassette.

**Cassette movement**

Some of the gene cassettes listed in Table 1 have been found in integrons of more than one class and it seems reasonable to conclude that all cassettes can be integrated into each of the three known integron types. However, to date all experimental work on cassette movement described below has been carried out with class 1 integrons.

To date, only one protein, the integron-encoded integrase IntI, has been shown to be required for the movement of gene cassettes. IntI is believed to be a member of the λ integrase family on the basis of similarities between the conserved domains shared by members of this family of site-specific recombinases (Ouellette & Roy, 1987; Stokes & Hall, 1989; Abremski & Hoess, 1992). IntI has been shown to catalyse both integrative and excisive recombination events in *in vivo* experimental systems (Martinez & de la Cruz, 1990; Hall et al., 1991; Recchia et al., 1994). IntI has also been shown to be essential for the excision, integration and reassortment of gene cassettes (Collis & Hall, 1992a, b; Collis et al., 1993). The specific recombination sites known to be recognized by IntI are the cassette-associated 59-base elements (Martinez & de la Cruz, 1990; Hall et al., 1991) and the integron-determined integration site, *attl*, which is largely contained within the 5’-CS and is not obviously related to the 59-base elements (Recchia et al., 1994) (Fig. 7).

Cassette integration and excision occurs via a conservative site-specific recombination mechanism (Fig. 2). In its simplest form, cassette integration involves a single site-specific recombination event between the 59-base element of a free circular cassette and the integron *attl* site (Collis et al., 1993). Cassette excision is the reverse of integration, though in this case the sites involved may be either *attl* and a 59-base element or two 59-base elements (Hall & Collis, 1992a, b). IntI-mediated excision of a cassette from an integron regenerates the circularized form of the cassette (Collis & Hall, 1992b). Other routes involving two site-specific recombination events can also lead to the reciprocal acquisition and loss of cassettes from the arrays found in integrons (see Hall & Collis, 1995), and these routes may even predominate under normal circumstances. When free circular cassettes were integrated experimentally into integrons containing both *attl* and one or more cassettes and hence one or more 59-base elements, a strong preference for insertion at *attl* was observed (Collis et al., 1993). In other experimental systems this preference has not been observed. Rather, events involving two 59-base element recombination sites have generally predominated (Martinez & de la Cruz, 1990; Hall et al., 1991; Collis & Hall, 1992a). While this type of event is clearly important to effect the excision of any cassette that is not located in the first position (i.e. adjacent to the 5’-CS and thus *attl*) in a cassette array, the reasons for the different site preferences in different experimental systems may reflect constraints intrinsic to these systems and this requires further investigation.

Occasionally, recombination can occur between a 59-base element and a secondary (or non-specific) site (Francia et al., 1993; Recchia et al., 1994; Recchia & Hall, 1995) and lead to the integration of a cassette at the non-specific location (Fig. 2). This type of event, though it occurs at very low frequencies (Francia et al., 1993; Recchia et al., 1994) is likely to be important in the dissemination of

![Fig. 7. Comparison of the sequences of *attl* and a representative 59-base element (*aadB*). For *aadB*, solid lines indicate the extent of the 59-base element determined by consensus (60 bases). The *attl* site is from class 1 integrons, and the solid and dashed lines indicate the estimated minimum and maximum extent of sequences required for full site activity (40 and 70 bases, respectively). Horizontal arrows denote regions of inverted repeatedness in each site. The core sites are shown in bold and are boxed, and vertical arrows show the position of the recombination crossover. The inverse core site (GCCTAAC) of the *aadB* 59-base element is also boxed. No inverse core site has as yet been identified in the *attl* sequence.](image-url)
cassette-associated genes and in the evolution of plasmid and bacterial genomes. Excision of cassettes from secondary sites is unlikely to occur as these cassettes are not flanked by two fully active recombination sites (Recchia & Hall, 1995), and integration of a cassette at a secondary site thus leads to the permanent acquisition of a new gene.

**Expression of cassette genes**

The known gene cassettes do not generally include promoter signals. Indeed in many cases less than 10 bases separate the first in-frame initiation codon from the 5' boundary of the cassette. Exceptions are the *cmlA* cassette that includes both a promoter and translational attenuation signals (Stokes & Hall, 1991) and is described in more detail below, and the *qacE* cassette which includes a weak promoter (Guerineau *et al.*, 1990). Cassettes are normally found integrated at *attI* in only one of the two possible orientations, and this orientation-specificity is presumably determined by some feature of the recombination sites. As a consequence, the genes are all in the same orientation, which allows their expression from a promoter, *P*ₐₕₜ, in the integron 5'-CS (see Fig. 2). In class 1 integrons, four variants of this promoter are known (Stokes & Hall, 1989; Bunny *et al.*, 1995). These variants differ in strength over at least a 20-fold range (Lévesque *et al.*, 1994; Collis & Hall, 1995), and the strongest has an activity equivalent to or greater than that of the tac promoter (Lévesque *et al.*, 1994; C. Tomaras, H. W. Stokes & R. M. Hall, unpublished observations). In a few integron variants a second promoter, *P*₂, has been created by the insertion of three bases between otherwise poorly spaced −35 and −10 signals. *P*₂ is a strong promoter (Lévesque *et al.*, 1994) and increases the level of resistance expressed when the weakest version of *P*ₐₕₜ is present (Collis & Hall, 1995).

Integration of multiple cassettes at *attI* creates new operons, as all cassette-encoded genes, whether in the first cassette or in promoter-distal cassettes, are dependent on *P*ₐₕₜ for their expression. Recent studies have demonstrated the existence of long transcripts covering several cassettes (Collis & Hall, 1995). However, the majority of transcripts are shorter. All originate at *P*ₐₕₜ but end prematurely at discrete points that appear to correspond to the ends of cassettes. These observations raise the possibility that the 59-base elements, which generally include inverted repeats, either act as transcription terminators or as processing signals for endonucleolytic cleavage of transcripts (Collis & Hall, 1995). This effect leads to differences in the levels of transcripts from promoter proximal and promoter distal genes. These differences reflect differences in the level of resistance conferred by the same cassette gene when its cassette is first or located downstream of other cassettes, and in general the first cassette is expressed at the highest level (Collis & Hall, 1995).

When cassettes are integrated at secondary sites, expression is dependent on the presence of a suitably oriented promoter in the target. In the case of the *aadB* cassette found integrated at a secondary site in a plasmid equivalent to RSF1010, a promoter is present and gentamicin resistance is expressed (Recchia & Hall, 1995). However, it is clearly possible that in other cases the cassette gene could remain silent if no promoter were available.

Translational initiation signals, i.e. potential ribosome-binding sites (RBSSs), have been identified preceding several cassette-associated genes and are located within the cassette boundaries. However, in other instances, particularly when the first in-frame initiation codon is located close to the cassette boundary, no obvious RBSS is present. In some of these cases further potential in-frame initiation codons are present but whether these are used is not known as, in general, the sequences of the N-terminus of the protein products have not been determined. However, for the *aadA2* gene where the first in-frame initiation codon, an ATG located 10–12 bp from the cassette boundary, is not preceded by an RBSS, evidence consistent with the use of a GTG initiation codon 12 bp downstream has been reported (Bito & Susani, 1994) and this GTG is preceded by an RBSS. For several cassette genes, the gene coding region is placed such that when the cassette is in the first position in an integrated array the gene is fused in-frame with an ORF of 69 codons in the 5'-CS (see e.g. Cameron *et al.*, 1986), and this could affect the length of the polypeptide product. However, it is unlikely that translation of these genes is initiated within the 5'-CS as RBSSs are not found in association with any of the potential initiation codons within this ORF. However, in certain integrons a duplication of 19 bp of the 5'-CS sequence alters the sequence such that translation can potentially be initiated within the 5'-CS at an ATG that is preceded by an RBSS leading to a short N-terminal extension of the protein encoded by the first cassette (Wohleben *et al.*, 1991), and the product of the *aacC1* gene has an N-terminal sequence that indicates that translation commences at this position. However, the *aacC1* gene is also expressed when the cassette is located downstream in a cassette array, indicating that a suitable initiation site is also present within the cassette (Collis & Hall, 1992a). Furthermore, the level of expression of the *aacA4* and *aadA2* genes, measured by the level of antibiotic resistance conferred, does not appear to differ when the cassette abuts a 5'-CS that does or does not contain the 19 bp duplication (C. M. Collis & R. M. Hall, unpublished observations).

**Expression of the *cmlA* gene**

The *cmlA* gene cassette found in the integron In4 of Tn1696 is unusual in that it includes both a promoter and translational attenuation signals (Stokes & Hall, 1991), and the promoter has been shown to be responsible for the expression of the *cmlA* gene (Stokes & Hall, 1991; Bissonnette *et al.*, 1991). Expression of *cmlA* is inducible by sub-inhibitory levels of chloramphenicol and the region upstream of *cmlA* includes a leader peptide and inverted repeats capable of forming alternate stem–loop structures (Stokes & Hall, 1991). These features are similar to those found upstream of the inducible *cat* and *ermC* genes of Gram-positive bacteria which are known to
be regulated by translational attenuation (for reviews see Lovett, 1990; Weisblum, 1995).

Origins of gene cassettes

The origins of the cassette-associated genes and of the cassettes themselves are fascinating questions. Based on the features of gene cassettes, namely that they contain only a single gene with little flanking non-coding sequences and a unique 59-base element, we have previously speculated that cassettes may originate from transcripts that are converted to DNA (Hall et al., 1991). This would presumably involve the action of an as-yet-unidentified reverse transcriptase. With this scenario, the cassette-associated recombination sites (59-base elements) could either be part of the processed transcript, e.g. a transcription terminator, or added to it prior to its conversion to DNA. Alternatively, the 59-base element could be acquired at a later stage.

Though there is no real evidence to support this proposition, the burgeoning of available DNA sequences for both cassettes and other resistance genes, and for the genomes of plasmids, transposons and bacteria, now permits the examination of some questions that may ultimately shed light on the origins of gene cassettes. For example, several sets of cassette genes are quite closely related, with 60–95% identity between individual pairs of DNA sequences within the coding regions. In some of these groups, for example aadA1 and aadA2, the same level of relatedness extends to the cassette boundaries, thus including the 59-base elements, and these cassettes may have diverged from a common ancestral cassette. However, in other cases quite different 59-base elements are associated with each of the genes in the set. For example, the dfrA1, dfrA5 and dfrA7 genes are between 62 and 67% identical but the 59-base elements are 95, 87 and 134 bases long, respectively, and are not obviously related (Sundström et al., 1993) except that they all contain the features common to 59-base elements and the first and last 20 bases all show similarity to the consensus for these regions (Hall et al., 1991; Collis & Hall, 1992b). The catB2 and catB3 cassettes also contain quite different 59-base elements (Bunny et al., 1995). This is consistent with a separate origin for the gene and the 59-base element in each cassette. The fact that quite closely related 59-base elements are also found associated with unrelated genes supports this conclusion. The largest group of 59-base elements includes those associated with the aadA genes, aadB, catB3, orfD and aacA (IIia) genes and each of these 59-base elements differs from the consensus for this group (Hall et al., 1991) by only a few bases. Other cases in which closely related 59-base elements are associated with different genes have also been found; for example, the aacA4 and aacA (orfB) genes encode proteins belonging to distinct protein families but are associated with very similar 59-base elements (see Hall et al., 1991).

If cassettes are indeed constructed from separate pools of genes and 59-base elements, then when potential progenitors of these genes, e.g. chromosomal equivalents of cassette genes, are identified they would not be expected to be associated with downstream sequences related to the 59-base elements found in the final cassette. Only one case of potential progenitors for cassette-associated genes has been identified to date. These are an ORF (of unknown function) in the chromosome of Pseudomonas aeruginosa which shares 67.5 and 69.5% identity with the catB2 and catB3 genes, respectively, and the catB1 gene in the chromosome of Agrobacterium tumefaciens which shares 66.0% and 64.8% identity with catB2 and catB3, respectively (Bunny et al., 1995). In both cases the similarity is confined to within the gene coding regions.

Gene cassettes are not transposons

Though gene cassettes are mobile elements, they are distinct from transposons in a number of ways. Firstly, they do not include any genes encoding proteins that catalyse their movement. The integrase is encoded by the integron. Secondly, cassettes are not bounded by inverted repeats, nor are they flanked by a duplication of target sequences. Thirdly, the mechanism used for mobilization of cassettes is distinct from transposition, as it involves conservative site-specific recombination. Gene cassettes thus form a family of mobile elements that are more akin to integrating phage (e.g. λ) and plasmids than to transposons. However, cassettes also form a distinct subgroup of the family of mobile elements that utilize conservative site-specific recombination for movement. Most other mobile elements of this family encode the site-specific recombinase (integrase) responsible for their movement whereas gene cassettes do not. In addition, integrating phage encode both phage proteins and replication functions, integrating plasmids encode replication functions, and the conjugative transposons (which integrate by conservative site-specific recombination and thus are not strictly speaking transposons) encode conjugation functions. In contrast, gene cassettes include only a single gene.

Further distinguishing features of the gene-cassette–integron system are that all known gene cassettes are recognized by the same integrase, despite differences in the precise sequence of their 59-base element recombination sites, and that three known integrases, each with different sequences, appear to be able to recognize the same family of cassettes.

Evolutionary implications

The gene-cassette–integron system is a multi-component system that functions as a natural cloning and expression system and is able to create arrays of genes of enormous diversity. Over 40 gene cassettes and three quite distinct classes of integron have so far been identified and it seems likely that further cassettes and further integrons will be identified in the future. That the majority of gene cassettes identified thus far encode antibiotic-resistance determinants is most likely a reflection of intense selective pressure and the fact that they all originate from clinical isolates. Granted the diversity of the resistance genes, it seems reasonable to assume that genes other than anti-
biotic-resistance genes may also be contained in cassettes. If this is so, the gene-cassette–integron system may be of much more widespread significance, and provide a general mechanism for the dissemination of modular gene units. The notion that genomes are composed of discrete modules (Shapiro, 1993) is an attractive one, and gene cassettes provide clear evidence for the existence of a modular gene-packaging system. This system is thus likely to be an important factor not only in facilitating the spread of antibiotic-resistance genes, but also in the natural engineering of bacterial genomes. The ability of integrons to create new operons (cassette arrays) that can then be fixed in place by deletion of the 5' base elements has obvious potential for the evolution of stable new operons in plasmid and bacterial genomes. The integration and consequent fixation of gene cassettes at secondary recombination sites is also a potential general mechanism for moving genes to new locations.

Acknowledgements

We wish to thank Paul Roy for critical reading of the manuscript and Lars Sundström, Paul Roy and Hilary-Kay Young for making available unpublished data. G.D.R. is supported by a CSIRO Postgraduate Scholarship.

References


Cameron, F. H., Groot Obbink, D. J., Ackerman, V. P. & Hall, R. M. (1986). Nucleotide sequence of the AAD(2') aminoglycoside adenyltransferase determinant aadB. Evolutionary relationship of this region with those surrounding aadA in R538–1 and dbfR in R388. Nucleic Acids Res 14, 8625–8635.


Mobile gene cassettes


