# The Insertion Sequences of *Anabaena* sp. Strain PCC 7120 and Their Effects on Its Open Reading Frames<sup>⊽</sup>†

C. Peter Wolk,<sup>1,2\*</sup> Sigal Lechno-Yossef,<sup>1</sup>‡ and Karin M. Jäger<sup>1</sup>§

MSU-DOE Plant Research Laboratory<sup>1</sup> and Department of Plant Biology,<sup>2</sup> Michigan State University, East Lansing, Michigan 48824

Received 22 April 2010/Accepted 14 July 2010

Anabaena sp. strain PCC 7120, widely studied, has 145 annotated transposase genes that are part of transposable elements called insertion sequences (ISs). To determine the entirety of the ISs, we aligned transposase genes and their flanking regions; identified the ISs' possible terminal inverted repeats, usually flanked by direct repeats; and compared IS-interrupted sequences with homologous sequences. We thereby determined both ends of 87 ISs bearing 110 transposase genes in eight IS families (http://www-is.biotoul.fr/) and in a cluster of unclassified ISs, and of hitherto unknown miniature inverted-repeat transposable elements. Open reading frames were then identified to which ISs contributed and others—some encoding proteins of predictable function, including protein kinases, and restriction endonucleases—that were interrupted by ISs. *Anabaena* sp. ISs were often more closely related to exogenous than to other endogenous ISs, suggesting that numerous variant ISs were not degraded within PCC 7120 but transferred from without. This observation leads to the expectation that further sequencing projects will extend this and similar analyses. We also propose an adaptive role for poly(A) sequences in ISs.

Insertion sequences (ISs) are transposable elements found in prokaryotic and eukaryotic genomes (17). A fully functional bacterial IS comprises one or more transposase genes, ends that are often inverted repeats (IRs), and, between the transposase genes and the ends, sequences termed linkers (32). Diverse bacterial ISs have been classified, and a searchable database of ISs has been constructed (ISfinder [http://www-is .biotoul.fr/]) (28). Miniature inverted-repeat transposable elements (MITEs) and even smaller mobile elements lack their own transposases and are also found in *Anabaena* spp. (11, 12, 33).

Anabaena sp. strain PCC 7120 (also known as Nostoc sp. [25], here denoted Anabaena sp.) is widely used to study the patterned differentiation of dinitrogen-fixing cells called heterocysts. Transposition of ISs in Anabaena sp. has been documented (1, 7–9). We earlier reported, with few details, three genes that are intercepted by ISs in Anabaena sp. (23). We here describe the approach more extensively, organize the ISs of Anabaena sp., and present our efforts to identify Anabaena sp. open reading frames (ORFs) interrupted or contributed to by ISs.

#### MATERIALS AND METHODS

**Software.** Sequences most similar to the sequence of a particular transposase were identified by BLAST search (2), often using BioBike (http://biobike.csbc .vcu.edu/) to obtain flanking sequences simultaneously. To determine how far in each direction an IS extends and what DNA, if any, was duplicated upon its

\* Corresponding author. Mailing address: MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI 48824. Phone: (517) 353-2049. Fax: (517) 353-9168. E-mail: wolk@msu.edu.

<sup>v</sup> Published ahead of print on 23 July 2010.

insertion, forming a direct repeat (DR) (14), the transposase genes and their flanking sequences were aligned by ClustalW in BioEdit (http://www.mbio.ncsu .edu/BioEdit/BioEdit.html) and/or manually. Identities or near identities and the sudden cessation of identity were sought and were often found at common distances from the ends of the transposase ORFs. In the figures with this report, we often juxtapose the sequences of ORFs with the sequences of longer regions, abbreviated "rn," that contain those ORFs so as to distinguish the two. Determination of the ends of ISs (Table 1, penultimate column) was often facilitated by the presence of IRs, frequently flanked by DRs, at the ends of many ISs or by the presence of possible target sites.

Assignment of ISs to known IS families used the search or BLAST functions of ISfinder (28). The bl2seq function of NCBI (http://blast.ncbi.nlm.nih.gov/Blast .cgi) was used with default settings to provide a measure of the expect (E) value that the left (L) end was significantly similar to an inversion of the right (R) end. Phylogenetic analyses of amino acid sequences, conducted in MEGA4 (30), were used to infer an evolutionary history by use of the neighbor-joining method (26). When more than one ORF was present in an insertion sequence, those ORFs were catenated and used as one sequence for the phylogenetic reconstruction. Evolutionary distances, in units of number of amino acid substitutions per site, were computed using a Dayhoff matrix (27). Positions containing alignment gaps and missing data were eliminated using the pairwise deletion option of the software.

**Nomenclature.** Although ISs are normally named IS followed by one or more italicized digits, e.g., IS1594 or IS5, not all of the ISs discussed have such names or can, with assurance, be assigned such names by comparison to known ISs. Often, two or more neighboring ORFs of *Anabaena* sp. are annotated as encoding transposases that may be part of the same IS. Because particular members of a set of ISs were often considered, an IS that bears a particular ORF or ORFs is often referred to as IS(*that ORF*) or IS(*those ORFs*), e.g., IS(*alt7001*). A prime appended to a transposase ORF indicates that it has been interrupted.

#### RESULTS

The ISs of Anabaena sp. The frequency of annotated transposase genes per megabase pair of genome varies widely within the cyanobacteria whose genomes have been sequenced, from none (some marine species of *Synechococcus* and *Prochlorococcus*) to 104 (*Microcystis aeruginosa* strain NIES-843). Many strains have between 10 (*Nostoc punctiforme* strain ATCC 29133) and 32 per Mb (*Synechococcus* sp. strain PCC 6803). Table 1 introduces the annotated transposase genes of

<sup>‡</sup> Present address: DOE Great Lakes Bioenergy Research Center, Michigan State University, E. Lansing, MI 48824.

<sup>§</sup> Present address: Radolfzeller Str. 78, D-78467, Konstanz, Germany.

<sup>†</sup> Supplemental material for this article may be found at http://jb.asm.org/.

	TABLE 1. Clustered ISs	of Anabaena sp.: h	low their ends w	ere identified, se	lected chara	acteristics, and re	elated supplemental	figures	
Family/group <sup>a</sup>	CL and/or Anabaena sp. PCC 7120 CL members <sup>c</sup>	Other members of the $CL^d$	L end (excluding variants) <sup>e</sup>	Inverted R end (excluding variants) <sup>f</sup>	E value of IRs <sup>g</sup>	Typical DR (no. of bp)	Tentative target and IS insertion point $()^i$	Basis for identification of ends <sup>i</sup>	Supplemental figure reference
IS4/IS50	IS(abr1332), 5 MITTEs	4 <i>Sy</i> 7002	<u>CTAC</u> G <u>GTGTA</u> <u>CACAC AAGT</u> C <u>CAA</u> GT AAAGT	<u>CTAC</u> C <u>GTGTA</u> <u>CACAC</u> <u>AAGTA</u> <u>CTA</u> TC TGCAA	5E-9	10 or 11	GC-rich	A, IRs, DRs	S1 CL 1
IS4/IS4Sa	1S(alr5204)	2 Av, 7 Np	$\begin{array}{c} \underline{C}\underline{A}\underline{G}\underline{A}\underline{A}\\ \underline{G}\underline{A}\underline{A}T\underline{G}\\ \underline{G}\underline{A}\underline{A}T\underline{G}\\ \underline{A}\underline{A}\underline{A}\underline{A}\\ \underline{A}\underline{A}\underline{A}\underline{A}\\ \underline{G}\underline{A}\underline{A}\underline{A}\\ \underline{G}\underline{A}\underline{A}\underline{A}\\ \underline{A}\\ \underline{A}\underline{A}GC\\ \underline{G}\underline{A}\underline{A}\underline{A}\\ \underline{A}\\ \underline{A}\underline{A}GC\\ \end{array}$	<u>Caaaa</u> atgt <u>t</u> <u>Gaaag</u> c <u>tgat</u> <u>Acaaa</u> at <u>t</u> t ac <u>ata</u> g <u>t</u> tag	0.058	Imperfect 9	AT-rich	A, IRs	S1 CL 2
IS4/ISPepr1	1S(all7115)	7 Np	<u>CAATA</u> <u>CCT</u> TA <u>GCCAA</u> <u>AATAA</u> <u>GAGCA <u>TAAA</u>G AGGTA <u>G</u>GGCG</u>	CAATA CCTCT GCCAA ATTAC GAGGG <u>T</u> TCAA CACCC <u>G</u> TAGA	0.015	9	Usually AT-rich	A, IRs, DRs, Table 2	S1 CL 3
IS5/IS1031	CL 1: IS(all0016.15), IS(all2693- 92), IS(ah3610-11), IS(all4400, all4399), IS(ah4438-39), IS(all4817-16), IS(ah5157-58), IS(al17902, all7001)		<u>GAGG CTATT</u> TATAA <u>AGTAA</u> <u>ATCTA AAGGA</u> <u>GA</u> GCT AT <u>C</u> AG	GAGA CCATT TATAA AGTAA ATCTT TAGAC GACTA GACGA	6E-8	ĸ	TWA (one TCA)	A, IRs, DRs, Table 2	S2 CL 1
IS5/IS1031	CL 2: IS( <i>ah</i> 7025)	Am, 3 Np	GAGG CTGTT TGAAA ACTAG GGGAT GT <u>T</u> GT AAAAA AACT <u>C</u> CC <u>TCG</u> G <u>T</u> AT <u>A</u>	GAGG ATGTT TGAAA <u>AGT</u> TA TA <u>G</u> GG AGTCA <u>AAA</u> TT <u>AAGCC</u> AA <u>ATCG</u> C <u>T</u> TC <u>A</u>	6E6	e	TWA	A, IRs, DRs	S2 CL 2
IS110/(-)	<b>IS1594</b> : IS( <i>alt</i> /0249), IS( <i>all</i> 0306), IS( <i>all</i> 0752), IS( <i>all</i> 1099), IS( <i>alt</i> 0752), IS( <i>all</i> 1986), IS( <i>alt</i> 12065), IS( <i>al</i> 13757), IS( <i>al</i> 13754), IS( <i>all</i> 3755), IS( <i>al</i> 13754), IS( <i>al</i> 14756)	dN	T <u>G</u> TAT A <u>T</u> TAA AA <u>GAA</u> GTGGT <u>AG</u> ACC GTCGC	A <u>G</u> CGA C <u>T</u> GTC TT <u>GAA</u> AGTCA <u>AG</u> CGA TCGTT	NSS	0	CCT/AC, CC/TAC, or C/CTAC	A, K (rRNA), Fig. 1B	S3
IS200-IS605/IS1341	<b>IS891</b> : IS(all3986), IS(ab4104), IS(all5207), IS(ab7228), IS(ab7231), IS(all8010)	Ns, 3 Lyn	GA <u>G</u> C <u>C</u> G <u>T</u> GAA G <u>C</u> GTA AAGCC CCCGT ATTTT	TT <u>GAC</u> A <u>T</u> CCT C <u>C</u> CCC GTTTA GAAAA CGGGG	0.032	0	TTAC/	K (see text), TGTCAA at R terminus, Table 2	S4 CL 1
IS200-IS605/IS1341	<b>IS891-related CL 2:</b> IS( <i>all0315-14</i> ), IS( <i>alr1157</i> ), IS( <i>all4465</i> ), IS( <i>alr7325</i> )	Ąv	CAAGA <u>A</u> ACTG GGT <u>C</u> T <u>AAA</u> GC CCC <u>G</u> T CCTTG	TTGAC <u>A</u> CTCT CCG <u>C</u> C CT <u>A</u> TA AGT <u>G</u> C GGAGA	NSS	0	TTAC/	A, TGTCAA at R ter- minus, Table 2	S4 CL 2
IS200-IS605/IS1341	18891-related CL 3: IS(all2167, alr2168), partial IS(all1608)	9 Te	CAAAA GAATG GGATA CAAGC CCC <u>G</u> T CGTTC T <u>A</u> GGA CG <u>G</u> CT	TTGAC ATACT CACCG ACCTA AAG <u>G</u> T GCGGT G <u>A</u> TTC TT <u>G</u> AC	NSS	0	TGAC/	A	S4 CL 3
IS200-IS605/IS1341	18891-related CL 4: IS(alr1531) (left end unclear)	Av, 7 Te	<u>TGG</u> TA <u>A</u> AATG TGAGG TATGG <u>AAA</u> AA <u>GC</u> CTA CCGCT ACCGA	<u>TTGAC ATCCT</u> CACCG CCCT <u>G</u> <u>AAAGT GCGGT</u> GATTC CTAAG	NSS	Unclear	Unclear	A, TGTCAA at R terminus, L terminus unclear	S4 CL 4
IS200-IS605/IS1341, (-)	<b>IS891-related CL 5:</b> IS(all7148, alr7149), IS(all7008, alr7009)	Av, Cy7424	АGTTT СТСАА АААТА ТАТТG АТGTT <u>АG</u> АС <u>G</u>	TTGAC ACTCT CGCCG CTAAC CGCAA <u>AG</u> CA <u>G</u>	NSS	0	TTAC?	A, TGTCAA at R terminus	S4 CL 5
IS200-IS605/IS608	IS891-related CL 6: IS(all3371, ah:3372), IS(all7085, ah7086) (approximately, respectively, ISN'92 and ISN'973 of ISfinder)	2 Np	GA <u>G</u> TC GTGA <u>T</u> G <u>C</u> GTA AAGCC CCC <u>AA</u> TTAT <u>G</u>	TT <u>G</u> AG CCAC <u>T</u> C <u>C</u> CCC GTTTT GAA <u>AA</u> CGGG <u>G</u>	NSS	0	TTAC	A, TCAA at R terminus	S4 CL 6

5290

J. BACTERIOL.

IS200-IS605 <sup>b</sup>	IS(ah-1015), IS(ah-4734)	Cy8801, Cy0110, Mc; ISLjo5_ a1	GTAG <b>G GTGGG CAATG CCCAC</b> CAAAA ATATT ATGTA AAAAT	GTAGG GTGGG CAITTG CCCAC СААТТ АТСТС АТТАТ GTAGT	4E-9	σ		A, IRs	S4 CL 7
IS630/(-)	IS895: IS(ah0552-53), IS(ah1726- 27), IS(ah1853-54), IS(ah1972- 71), IS(ah12067-66), IS(ah12073- 74), IS(ah12067-66), IS(ah14868-67)		TAGGA ATCCT ATTTG ATTTG TGAAC AAGAC CAAGA	TAGGA ATCCT ATTTG ATTTT TGAAT AAGT CCGTA	8E-10	7	TA	A, IRs	S5 CL 1
IS630/(-)	IS895-related CL 2: IS(as/1992), IS(asr3082) (diverges from others close to R end)	Av, Np, 2 Cw, 4 Am	<u>ACCAA</u> TTTAA <u>AT</u> TAG AGAC <u>A</u> <u>GGGCA</u> <u>GAT</u> GA <u>GGTA</u> A	<u>Accaa</u> A <u>ttaa</u> <u>At</u> ggt gttt <u>a</u> <u>Gggca</u> <u>Gatt</u> ag <u>Ggca</u> t	0.003	7	A <u>TA</u> T	A, IRs	S5 CL 2
IS630/(-)	<b>IS895-related CL 3:</b> IS( <i>alr018-19</i> ), IS( <i>al10363-62</i> ), IS( <i>alr1858-59</i> )	См, Np, Су7425	TAGCG TTTAC CAGTA TAATG AAGTA CACTA <u>ATTA</u> AATAA	TAGCG TTTCT CAGTC TGGTG AAGTA CAGTA <u>AAGTA</u> AATGA	0.004	0	TA	A, IRs, Table 2	S5 CL 3
IS630/(-)	IS895-related CL 4: IS(ab527-28)	Av, 7 Ns	<u>AGTAG</u> <u>GTAGG</u> CAC <u>GA</u> <u>AAA</u> AA CCAAA <u>CTATG</u> CCAAA <u>CTATG</u> TG <u>AAG</u> <u>ATAAG</u> TAAAG ACGG <u>A</u> <u>GAATA</u> <u>A</u> ATC <u>T</u>	AGTAG <u>GTGGG</u> TGG <u>GA</u> <u>AAAG</u> T CCCAA <u>GTATG</u> <u>TAACG AAA</u> CA <u>TTAAG</u> <u>AAA</u> CA <u>TTAAG</u> TAAT <u>A</u> <u>GAAT</u> T <u>A</u> GAG <u>T</u>	0.011	0	<u>VI</u>	A, IRs	S5 CL 4
IS630/(-)	IS895-related CL 5: IS(all7564-63)	17 <i>Cw</i>	<u>GTACA</u> GG <u>TCG</u> <u>GCGTA</u> <u>AATAA</u> A <u>CAGA</u> <u>CCAT</u> A	<u>GTACA</u> CC <u>TCG</u> <u>GCGTA</u> <u>AATCA</u> <u>GCAGA</u> <u>CCAT</u> T	2E-6	7	CTAG	A, IRs, Table 2 (unchanged reading frame)	S5 CL 5
IS892 (unclassified) [aka <i>IS(al17268)]<sup>k</sup></i>	IS892: IS(al7005-04), IS(al17106- 05), IS(al7112-11), IS(al17178- 77), IS(al17303-02), IS(alr7323), IS(al7376-75), IS(alr8510), IS(alr8566, asv8561, alr85202)		CTAGC GTGGC AAAAC TTACT AGAGA GGAGC <u>A</u> GAGA TCCTG	<u>CTAGC</u> <u>GTGGC</u> <u>AAAAC</u> <u>TTACT</u> <u>AGAGA</u> <u>AGA</u> CG <u>A</u> CTCT CTAGA	2E-12	ર્જ	AT-rich	A, IRs, DRs, Table 2	S6 CL 1
IS982/(-)	CL 1: IS(ah/099), IS(al/2664), IS(ah/2683), IS(ah/2694), IS(ah/3384), IS(all/3624) ISNsp1 [aka IS(ah/1569)]		<u>ACG</u> TG <u>ATGTG</u> <u>CGACT</u> <u>TATTG</u> <u>TTTCG</u> <u>TTACA</u> <u>CAATT</u> <u>GAGGT</u>	<u>ACG</u> CC <u>ATGTG</u> CGACT <u>TAATA</u> TTCTG <u>TAACA</u> AG <u>AT</u> C <u>GTCGA</u>	6E-5	mostly 6	AT-rich	A, IRs, DRs, Table 2	S7 CL 1
IS <i>982/(-)</i>	CL 2: IS(ast0588), IS(atr0590)	2 Gvi	<u>ACGTG</u> <u>AGTTC</u> <u>GACGG</u> <u>GTTAA</u> TTT <u>A</u> G GTG <u>A</u> A	<u>ACGTG</u> <u>AGTTC</u> <u>GACG</u> A AC <u>TAA</u> AAAAC AGC <u>A</u> G	2E-6	8 or less	AT-rich	A, IRs, DRs	S7 CL 2
IS982/(-)	CL 3: IS(all8559), IS(alr4082), IS(asr7385) (a fragment)	ISRmsp1	ATTTA GGGT <u>T</u> TGTGC GAGCC AACT <u>A</u> TTTGA	TACGC CTTA <u>T</u> GTGAA TTAAG CCGG <u>A</u> GGATG	SSN	1?	AT-rich	A, Table 2	S7 CL 3
ISAzo13/(-)	CL 1: IS(alr7562), IS(all8069)	ISStau6	<u>GAGAA</u> <u>CTGCA</u> <u>CAGAA</u> TGAT <u>T</u> <u>GATCC</u> TATGA TC <u>A</u> GA GAAAG	<u>GAGAA CTCCA</u> <u>CAAAAA</u> AAGA <u>T</u> GATCC AATAG CT <u>A</u> TG CTGGT	0.015	σ	AT-rich	A, IRs, DRs	S8 CL 1
ISA2013/(-)	CL 2: ISN <i>sp4</i> [aka IS( <i>alr8019</i> )], IS( <i>al12145</i> ) (truncated at its R end); IS( <i>asr7385</i> ) is an R-end fragment	8 Np	<u>AGG</u> CA <u>TCATG</u> <u>TAAAA</u> <u>ATAAC</u> <u>TTGAA</u> <u>CGATT</u> <u>TACCG</u> <u>AA<u>TAG</u> <u>TT</u>AGA</u>	<u>AGG</u> AG <u>TTATG</u> <u>TAAAA ATTAAC</u> C <u>TGAA</u> <u>CAATT</u> <u>AAGTG</u> <u>CCTAC</u> <u>TT</u> TGG	2E-6	m	TWA in AT-rich region	IRs, Table 2	S8 CL 2
								Continued on f	ollowing page

			TAB	LE 1—Continue	q				
Family/group <sup>a</sup>	CL and/or <i>Anabaena</i> sp. PCC 7120 CL members <sup>c</sup>	Other members of the $CL^d$	L end (excluding variants) <sup>e</sup>	Inverted R end (excluding variants) <sup>†</sup>	E value of IRs <sup>g</sup>	Typical DR (no. of bp)	Tentative target and IS insertion point $(/)^{i}$	Basis for identification of ends <sup>/</sup>	Supplemental figure reference
ISL3/(-)	CL 1: IS( <i>ah</i> 7609), IS( <i>ah</i> 7208), IS( <i>al</i> 17161), IS( <i>ai</i> 177305) (approximating IS4sp1), IS( <i>ah</i> 7349) (truncated), IS( <i>ah</i> 7350)	3 Np	<u>GGTTC TTTTCG</u> GA <u>TAT TTTAT</u> <u>GGAGA AAGCA</u> A <u>AAG TAACCA</u> A <u>AAG TAATG</u>	<u>GGTTC TTTGCC</u> CC <u>TGT TTTTAT</u> <u>GGAGA A</u> TTAA T <u>A</u> CTA A <u>A</u> GTO CC <u>A</u> G <u>T</u> <u>T</u> TAAT	$\mathrm{E}^{-4}$	Up to 8 perfect, often imperfect	AT-rich	A, IRs, DRs	S9 CL 1
(-)/£781	CL 2: IS(alr7386', alr7003', asr7006; alr7007), IS(alr8016-17)	2 Np	GGTTCTTGGCAACTTTTGGTGATCTTGGTTGGGGAAAGGCAGAAGCCAGGAGAAGCAAGGAGTCAGAAAGATTAAATTGA	GGTTC TTGGC   AACTT TTGGT   GATCT TGGTT   GGGGA AAGGC   AGAGG CAGA   GGGCA GAGGG   CAGAA GGCAG	5E-23	Up to 8 perfect, often imperfect	AT-rich	A, IRs, DRs	S9 CL 2
<sup><i>a</i></sup> ISs bearing the foll [CL] 1), <i>all7008</i> and <i>al</i>	lowing transposase ORFs cannot (yet) hr7009 (Fig. 4E and F), alr7153 (perha	) be excised computation aps in the IS607 family)	nally: IS5 family, al , all7158 (closely re	(12152; IS200/IS605 slated to the IS891	family, alr1531 transposase ge	, alr1685, alr2719, all ne), all7245, asl7246,	14675, all5207 (see Fig. , alr7329, all8070, alr80	S4 in the supplemental m 71; and unclassified by IS	aterial, cluster finder: alr1015

and *alr4734* (see the text and Fig. S4, CL 7); IS481 family, *all3630*; IS607 family, *ar7146*, *ab7147*, *ar7152*; IS630 family, *all9054*, *alr1926*, *ar3982* (see Fig. S5, CL 3); IS982 family, *all0588* (very short, but retains IR and DR; see Fig. S7, CL 2); *ar7385* (fragment), *alr4082*, and *all8559* (see Fig. S7, CL 3); IS1182 family, *alr9024*; ISAs1 family, *all8064*, *all8064*, *all8065* (see the text); IS4*z013* family, *all2145* (see Fig. S8, CL 2), *ar7385*; ISH3 family, *all7244*; not classified: *alr7163* (see the text). –, not assigned by ISfinder.

<sup>b</sup> Nunvar et al. (22).

<sup>c</sup> Cluster (CL) identifications are in boldface.

<sup>d</sup> The number of ORFs found in non-*Anabaena* strains that bear members of a cluster is given. Strains are abbreviated as follows: *Am*, *Acaryochloris marina* MBIC 11017; *Acma*, *A. marina* Acma49; *Av*, *Anabaena variabilis* ATCC 29413; Cy7425, Cyanothece sp. strain PCC 7425, Cw, Crocosphaera watsonii WH8501; Cy0110, Cy7424, and Cy8801, Cyanothece sp. strains PCC 0110, PCC 7424, and PCC 8801, respectively; *Gvi, Gloeobacter violaceus* strain PCC 7421; *Lyn, Lyngbya* sp. strain PCC 8105; *Mechococcus sp. strain* PCC 7420; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 8105; *Mechococcus sp. strain* PCC 7420; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 8105; *Mechococcus sp. strain* PCC 7420; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 8105; *Mechococcus sp. strain* PCC 7420; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 8105; *Mechococcus sp. strain* PCC 7420; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Synechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Mechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Mechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Mechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Mechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Mechococcus sp. strain* PCC 7421; *Lyn, Lyngbya* sp. strain PCC 7444; *Mechococcus sp. strain* PCC 7445; *Lyn, Lyngbya* sp. strain PCC 7445; *Mechococcus sp. strain* PCC 7455; *Mechococcu* 7002; *Te, Thermosynechococcus elongatus*. For other ISs, see ISfinder. <sup>*e*</sup> Identities to the inverted R end are underlined. Boldface indicates palindromic sequence. <sup>*J*</sup> Identities to the L end are underlined. Boldface indicates palindromic sequence.

<sup>8</sup> NSS, not significantly similar. <sup>h</sup> Confirming Cai (7, 8). <sup>h</sup> Duplicated sequence is underlined. <sup>j</sup> A, alignment; K, in known sequence. <sup>k</sup> aka, also known as.



FIG. 1. Phylogenetic relationships of amino acid sequences of transposase genes of *Anabaena* sp. ISs and their homologs in the following families are shown: IS5 (A), IS200-IS605 (IS1341, including IS891, and IS608) (B), IS630 (including IS895) (C), IS982 (D), ISAzo13 (E), and ISL3 (F). Percentages of replicate trees, greater than 50%, in which the associated transposases clustered together in the bootstrap test (1000 replicates) are shown above the branches. Each tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances, number of amino acid substitutions per site, that are used to infer a phylogenetic tree.

Anabaena sp., ca. 20 per Mb (http://genome.kazusa.or.jp /cyanobase/Anabaena; accessed 14 February 2010, with some changes resulting from this study), organized within families, subsets thereof called groups, and clusters of ORFs within the groups. Clusters obtained by phylogenetic analysis (Fig. 1) matched clusters that were obtained by bl2seq comparisons of sequences from the same IS family (data not shown). Anabaena sp. ORFs paired within ISs are paired also in Table 1, oriented upstream to downstream when the ORFs are parallel. Table 1 also presents characteristics of the ISs, with cross-references to the supplemental figures in which nucleotide and corresponding protein alignments are presented. Table 1 (footnote *a*) also identifies the ISs for which the two ends could not be identified with assurance.

**IS4 family.** Anabaena sp. ORFs *alr1332*, *alr5204*, and *all7115* encode transposases in the IS4 family (ISfinder). Alr1332 is extensively similar to approximately the first half of the predicted products of four *Synechococcus* sp. strain PCC 7002

ORFs whose ISs have extensive, perfect IRs (Fig. 2A). Although alr1332 has a similar L end and terminal sequence, no corresponding R end is found immediately after the end of its ORF. Instead, the alr1332 region continues as MITEa, one of five closely related elements (MITEa to MITEe) (Fig. 2B and C) whose L and inverted R ends closely resemble the L end of IS(alr1332) (Table 1). tBLASTn with the C-terminal portion of a Synechococcus sp. transposase as the query locates a likely R end of IS(alr1332) between bp 1139032 and 1138790 (Fig. 2D). However, whereas the MITEs and the PCC 7002 ISs mentioned are flanked by 10- or 11-bp DRs (Fig. 2B), the likely R end lacks a DR comparable to the sequence at the L flank of IS(*alr1332*) (Fig. 2A). tBLASTn found homologs of the MITEs' coding regions (Fig. 2C) in Nodularia spumigena CCY9414 (data not shown). A nucleotide sequence in Anabaena variabilis closely resembles the sequence found upon computational removal of MITEc and one copy of its DR (Fig. 2E).

The extensive homology of alr5204 and all7115 to ORFs of

Δ	40 50 60 70 80 90 1540 1550 1560 1570 1580
syn7002A2122 syn7002A2141 syn7002A2864 syn7002G0077 IS(alr1332) Possible re-	Image: construction in the initial initininitiali initial initial initial initial initial initi
В	
MITEa Asp158 MITEb Asp302 MITEc Asp357 MITEd Asp516 MITEe Asp pβ	13737ff ACAGCCGAGCG//CTACGGTGTACACAAGTCTTGGAATAACGCCTAACGCTTGGTTTCGTCGTCTAGCTCTAGGCTGTGCGATCGCTCTGGAAGCCTGTCA-TTGCGAGGAGGAACGACGAAG 16497ff <u>CCTGCAAGGAG//CTACGGTGTACACACAAGT</u> CTTGAAATAACGCCTGAGCTTGGTTTCGTCGTCTAGCTGTGGCGATCGCTCTGGAAGCCTGTCA-TTGCGAGGAGGAACGACGAAG 16734rn <u>CCTGCAAGGAC//CTACGGTGTACACACAAGT</u> CTTGAAATAACGCCTGACGCTGGGTTCGTCGTCTAGCTGTGGCGATCGCTCTGGAAGCCTGTCA-TTGCGAGGGAACGACGAAG 16739ff <u>CCTGCGAGCG//CTACGGTGTACACACAAGT</u> CTTGAAATAACGCCTGACGCTAGCGTTCGTCGTCTAGCTGTGCGATCGCTCTGGAAGCCTGTCA-TTGCGAGGGAACGACGACGAAG 16639ff <u>CCCCCCCGGG</u> // <u>CTACGGTGTACACACAAGT</u> CTTGAAATACGCCTAAGCTTGGTTTCGTCGTCTAGCTGTGGCGATCGCTCTGGAAGCCTGTCACTTGCGAGGAAGGA
130  . MITEA CAATCA MITEb CAATCO MITEC CAATCA MITEC CAATCA MITEC CAATCO	140     150     160     170     180     190     200     210     220     230     240     250     260
C MITEA MYTQVI MITEA MYTQVI MITEC MYTQVI MITEC MYTQVI MITEC MYTQVI	10   20   30   40   50   60              .sarvssrdpppnCHyerseviasscsfcDcFvVpprn-DrlpersHsLeLDDetKr*     AvrvsgvInPLNCHYERSbyRsevEviasscsfcDcFvVpprn-DrlpersHsLeLDDetKr*     .sarvssgtpppnCHyerseviasscsfcDcFvVpprn-DrlpersHsLeLDDetKr*     .sarvssgtpppnCHyerseviasscsfcDcFvVpprn-DrlpersHsLeLDDetKr*     .sarvssgtpppnCHyerseviasscsfcDcFvVpprn-DrlpersHsLeLDDetKr*     .sarvsgtpplnCHyerseviasscsfcDcFvVpprn-DrlpersHsLeLDDetKr*
D Right (R) en Possible R t	ad of Syn7002A2122: VVFSPEEQECLRQLAPTLQGKTQKQQNPHASASLSWATWLIGRLGGWSGYRSQSPPGIRTLWRGLYQFESIFHGWKLAQSTLVCTQ* erminus of Alr1332: LLFFHE*QQCLSTLAxLLPCKTRLQQNSYLPSATWIMQGLGGWSGYKSPKPPGITTWSRGLEQFESTFFGWKLALGKLVCTL*
L PCC 7120 seq Query: 961 Sbjct: 35770	uence: agoatcatagtgtttataacttctgctttgtctgcgacaccctctgcaaacttcct <u>cctgtaggaac</u> 1027 
	Query:     1017     cctgtaggaactggcaaataactagagagcaaagcagccacgcactcaatcgtgactgttt 1080       IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
A. variabili Query: 961 Sbjct: 11639	<i>is</i> sequence: agcatcatagtgtttataacttotgotttgtotgogacacocototgoaaacttoot <u>ootgtaggaac</u> tggcaaataaotagaggacaaggagcaaggocacgactoaatogtgactgottgttt 1080 
F Alr5204 "Ava" "Np"	10   20   30   40   50   60   70   80   90   100   110   120     viin*fpkivkdilkglpkndypvlnstlfefwlsyvldnsltsMRGLFARLNNTGFELDISTFSKANLHRSQKPFQAIVQKLNKLVQNKAENKLHNKYAIVSIDSTVITTSKLLWVL   10   10   120     VIINSFPKIVKDILKGLPKNDVFVLNSKLFFEFWLSVLDNSLTSMRGLFARLNNTGFELDISTFSKANLHRSQKPFQAIVQKLNKLVQNKAENKLHNKYAICFIDSTVITISKLLWVL   10   120     VIINSFPKIVKDILKGLPKNDVFVLNSKLFFEFWLSVLDNSLTSMRGLFARLNNTGFELDISTFSKANLHRSQKPFQAIVQKLNKLVQNKAENKLHNKYAICFIDSTVITISKLLWVL   10   120     VIINSFPKIVKDILKGLPKNDVFVLNSKLFFEFWLSVLDNSLTSMRGLFARLNNTGFELDISTFSKANLHRSQKPFQAIVQKLNKLVQNKAENKLHNKYAICFIDSTVITISKLLWVL   10   120
G	110 120 130 140 150 160 170
alr5204 rn alr5204 "ava" rn	TTTTGGGCTAGATATTACAATATTTTGATTTCACAGTGATTATAAATTAATT
"ava" "np" rn "np"	GTGATTATAAATTCATTTCCCAAAATTGTCAAAGATATCCT TTTTGGGCGGAAGATTAAAGTATATCTATTTAATAGTGATTAAAATTCATTTCCCCAAGATTGTCAAAGATATCTT GTGATTATAAATTCATTTCCCCAAGATTGTCAAAGATATCTT
н	1360 1370 1380 1390 1400 1410 1420 1430 1440
all7115 rn all7115 "np" rn	CCGTCTCTTAATCCTGACTCCGGCATTATTGATCTTAAAGCTCACCCAGAGTGGGTTTTCGATATGTCCGTGAAATGTTAAAAAATGCTC CCGTCTCCTTAATCCTGACTCCGGCATTATTGATCTTAAAGCTCACCCAGAGTGGTTTTCGATATGTCCGTGA- CCGTAAACATAATCCTGGTTCTGGCATTATTGACCTGAAAGCTTATT-ACCGTGGTTTTCGATATGTTCATGAAATGTTAAAAAATACTTC

"pp" CCGTAAACATAATCCTGGTCLTGGCATTATTGACCTGAAAGCTTATT-ACCGTGGTTTCGATATGTTCATGAAAGTTAAAAATACTTC

FIG. 2. Transposable elements in the IS4 family. Border regions of IS(*alr1332*) and related ISs from *Synechococcus* sp. strain PCC 7002 (A) and MITEs whose ends are similar to those of the ISs in panel A are shown (B). //, presumptive ends of ISs and MITEs; rn, region. The Syn7002 ISs extend from position 71 to position 1559, the MITEs extend from position 853 to 1559, and IS(*alr1332*) extends from position 71 to, possibly, the end of MITE Asp1583737 at position 1559. (C) When frameshifts (x) and stop codons (\*) are considered, the amino acid sequences of the MITEs appear very similar. (D) Predicted product of translation of what may be a repositioned fragment of the 3'-terminal portion of *alr1332*, compared with the predicted 3'-terminal product of *Synechococcus* sp. strain PCC 7002 ORF A2122. (E) BLASTn analysis using, as query, the "pretransposition" sequence of *Anabaena* sp. in the vicinity of MITEc. With *Anabaena* sp. as subject, one sees—spaced 224 bp apart—the two copies of the MITE's DR (underlined), whereas with *A. variabilis* as subject, one sees the unbroken region with the "empty site" without the MITE. (F to H) Single-base pair changes greatly affect the transposase amino acid sequence. (F) Alr5204 is truncated (the black, lowercase amino acid sequence lacks an in-frame start codon) compared with orthologs from *A. variabilis* (Ava) and *N. punctiforme* (Np). These proteins and their genes, *ava* and *np*, are specified in Fig. S1, clusters 2 (for F and G) and 3 (for H), in the supplemental material. (G) Nucleotide sequence comparison of *alr5204* and its orthologs shows that a single nucleotide substitution leads to creation of the stop codon shown in panel F. (H) The C-terminal truncation of *alr1715*, relative to np, results from an indel mutation at position 1398 that leads soon to a termination codon. Numbering of positions here and in other figures is consistent with that used in corresponding supplemental figures.

Α		40	50	60	1510	1520	1530	1540
						.		.
all0306	rn	ATTCATACTGACO	GCACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACTAACCG	TCAATTT
alr1212	2 rn	ATTCATACTGACO	CACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACTAACTT	GTATGCT
alr3571	. rn	ATTCATACTGACO	CACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACTAAAAA	ATATTTA
alr3636	5 rn	ATTCATACTGACO	CACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTTGCATCTC	AAAGAAG
all2065	5 rn	TCTCGCACTGACO	SCACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACAGTTTC	AATATTT
all3734	rn	TCTCACACTGACG	SCACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACTAGCGT	GTCAAGC
alr0249	rn	TATCGCACTGACO	CACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACCAAACC	ATCAATT
all0732	2 rn	TATCACACTGACC	GCACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACATTAAG	ATCACTT
all4756	5 rn	CTTGCAAGTAGCO	GCACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACTAACTA	AATAAGC
all3682	2 rn	CTATGGCGGAACA	ACACC//CTTG	TATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACAAAATT	SCCTTTT
all1099	) rn	GGTGGGCATTGCC	CACC//CTTG	FATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACAAATAC	TACATA
all1986	5 rn*	TACTACGGAACG	TCCC//CTTG	FATATTAA	GCTTGACTTTCAA	GACAGTCG//	CTACCGATTA	ATAAATT
rrn23Sb	rn	TACTACGGAACGC	TCCC				CTACCGATTA	ATAAATT
Npu4461	. rn	TCGTTCCTCAACO	CAAC//CTTG	TATCTTAG	GCTTGACTTTTAA	GACAGTCG//	CTACAGAACT	TAAGGTT
R								
D								
Query:	1	tcaacattat	tagtcgttagt	tttggtaggg <sup>.</sup>	tgcgtcagtatgaat	aatttctgag	gtatag 60	
Sbjct:	57978	381 tcaacattat	cagtcgttagt	tttggtaggg	tgcgtcagtatgaat	aatttctgag	gtatag 5797	822
Query:	61	ttaggttcta	atcacactgac	gcacc/ctac	attaagatcacttag	agcatttgag	gttegte 120	
				/				
Sbjct:	57978	321 ttaggtteta	atcacactgac	gcacc/ctac	attcagatca-ttgg	agcatttgag	gttcatc 579	7763
Query:	121	agttctatat	tcaacagecaa	acaqtcaaca	atcaac 156			
·· 4		11111111	11111111111					
Sbjct:	57977	62 agttctatat	tcaacagccaa	acagtcaaca	atcaac 5797727			
-		-	2					

FIG. 3. (A) Border regions of copies of IS1594. Presumptive ends (//) of the ISs could be moved to the right 1 bp (C//TTGTAT ... TCGC//T) or 2 bp (CT//TGTAT ... TCGCT//).\*, IS1594 copy within rrn23Sa. (B) A region of A. variabilis (Sbjct, subject) between ava4633 and ava4634 has great homology to both sides of the position (/) in which IS(all0732) is inserted in Anabaena sp. (Query).

N. punctiforme (see Fig. S1, clusters 2 and 3, in the supplemental material)-and of alr5204 also to ORFs of A. variabilisenabled the ends of their ISs to be identified. However, both ISs may be nonfunctional: Alr5204 lacks a long central region, and its N-terminal region (Fig. 2F and G) and the C terminus of All7115 (Fig. 2H) are truncated.

IS5 family, group IS1031. Whereas ISs of group IS1031 normally have a single ORF (ISfinder), ISs of cluster 1 bear paired, parallel, overlapping ORFs. IS(alr7025), the sole Anabaena sp. member of cluster 2, has only a single ORF, which closely resembles N. punctiforme ORFs NpR2130 and NpR4690 (Fig. 1A). The ISs of clusters 1 and 2 show extensive IRs and 3-bp DRs, usually TWA.

**IS110 family.** The only IS110 family ISs of *Anabaena* sp. are 12 identical copies of IS1594, each of which bears 1 of the 30 AvrII sites in its genome (5, 15). The copy reported in Gen-Bank (accession number AF047044; J. B. Bongianni and P. J. Lammers, unpublished data) is present within rrn23Sa, one of four otherwise identical 23S rRNA genes of Anabaena sp., enabling nearly unambiguous determination of the ends of the IS (Fig. 3A, legend) (24). That is, IS ends identified by alignment of the 12 copies (Fig. 3A) concur with the ends found by comparing the interrupted *rrn23Sa* sequence with the uninterrupted rrn23S sequences (Fig. 3A) and with an "empty site" sequence in A. variabilis that matches the sequences flanking IS(all0732) of Anabaena sp. (Fig. 3B).

Members of the IS200/IS605 family. This family, which includes IS891 and related ISs (Table 1), normally has one or two ORFs and neither IRs nor DRs (ISfinder). IS891 (GenBank accession number M24855.1) was reported first from Anabaena sp. strain M-131 within a previously established genetic sequence (4, 31), providing unambiguous identification of its ends. Figure 4A illustrates that, as in the insertion originally reported, TTAC precedes all insertions in Anabaena sp. of this cluster of IS891s. Therefore, these copies of IS891 may preferentially, perhaps obligatorily, target TTAC sites. With one exception, members of IS891-related cluster 2, whose transposases have diverged from those of IS891 (Fig. 1B), retain TTAC as a presumptive target site, but members of IS891related cluster 3 illustrate numerous instances of what appear to be targeted TGAC sites and an example of a TCAC (Fig. 4B). Figure 4C presents confirmatory evidence that TGAC lies immediately adjacent to the site of insertion of IS(all2168, all2167) at its L end. Other instances, e.g., as shown in Fig. 4D, require more evidence to distinguish what is within the L end of the IS and what may be the target.

The sequence TGTCAA is at the R end of the original IS891 and appears to represent the R end of many other IS891related ISs (Fig. 4A to D and G). Figure 4G presents an apparent exception. Homology of IS(all7008, alr7009) and other sequences in IS200-IS605 cluster 5 (Fig. 1B) is visualized by tBLASTx (Fig. 4E and F), permitting the assessment that the locus of insertion of IS(alr7386', etc.) (Fig. 4E and F, vertical arrowheads, and 5) is ca. 250 bp in from the end of IS(all7008, alr7009). [IS(alr7386', etc.) stands for IS(alr7386'alr7003'-asr7006, alr7007) lacking IS(all7002, all7001), IS-(all7005, all7004), and the direct repeats that their insertions occasioned.] The members of IS891-related clusters 3 and 6 (see Fig. S4 in the supplemental material), like those of cluster 5 (Fig. 4E and F), bear divergently oriented ORFs.

BLASTp shows weak similarity of Alr1015 to ISLio5, within the IS200/IS605 family. Alr1015 is also related to Alr4734 and has yet greater similarity to transposases of Cyanothece spp. and a Microcoleus sp. (Fig. 6A). Tyrosine

Α	30 40	50	60	70	80	90	1410	1420 14	30 1440	1450	1460
(IS891) all3986 rn alr4104 rn alr7228 rn alr7231 rn all8010 rn	CCCCCAGAAAC GCTTACTAAAA TTTTACGTATT ACCACTAAGAT GCTAATTTTCA	TTCTTCTAAGGT ACTGCGCGACAG ATTGTCCCGTTT TGCTTACTGACA GAAAAATTTTAC	tac/ttac//0 TGAG/TTAC//0 AATC/TTAC//0 GATC/TTAC//0 ATCC/TTAC//0 TAAT/TTAC//0	AGCCGTGAA AGCCGTGAA AGCCGTGAA AGCCGTGAA AGCCGTGAA	GCGTAAAGCCC GCGTAAAGCCC GCGTAAAGCCC GCGTAAAGCCC GCGTAAAGCCC GCGTAAAGCCC	CCCGTAT CCCGTAT CCCGTAT CCCGTAT CCCGTAT CCCGTAT	СТТТТСТААА СТТТТСТААА СТТТТСТААА СТТТТСТААА СТТТТСТААА СТТТТСТААА СТТТТСТААА	CGGGGGAGGAT CGGGGGAGGAT CGGGGGAGGAT CGGGGGAGGAT CGGGGGAGGAT CGGGGGAGGAT	GTCAA//ttgggc GTCAA//CCTATC GTCAA//AGAAAG GTCAA//AAAACT GTCAA//AACTAT GTCAA//AGTCTT	g GCCAAGCAT TGCTGTCAT TCGCGTCTT TTTGATATT CATACTTTA	TTGCAGGCA TAGAGGAGA TTCGCATCT TTTGAAGTA GTGCGTTTA
В											
Cluster 2	30	40 	50 6	50 	70 .	80 1340	1350	1360	1370 1	380 .	1390
all0315 rn alr1157 rn all4465 rn alr7325 rn ava1635 rn	CCGCAGGGTAC GACCGGACTAC GTAGTAATTTC CGCCCAGCGTA AACCTTTACCT	SCGGATAAATTT FAGCTAAAAAAT CGTACAATATAA SCTGATCGCAGC AAGTTCCCTAAC	/TTAC//CAAGA /TTAC//CAAAA /TTAC//CAAGA /TTAC//CAAGA /TGAC//CGAAA	A-ACTGGGTC A-AATGGGTC A-ACTGGGTC A-ACTGGGTC AGAATGGGAT	CTAAAGCCCCC CTAAAGCCCCC CTAAAGCCCCC CTAAAGCCCCC FACAAGCCCCC	TCGC2 TCGC2 TCCGC2 TCCGC2 TCCCC2 TCCCC2	ACTTAtagGGCC CCTTAtagGGCC ACTTAtagGGCC ACTTAtagGGCC CCTTAtagGGCC	GAGAGTGTCAA GAGAGTGTCAA GAGAGTGTCAA GAGAGTGTCAA GAGAGTGTCAA	//CTCAGTAGACT //AACTCATCTCT //CTATATTCCCT //AGTTTATTGCT //CAATTAAGTAG	AGGAAACGC CATCTCACG CATATCCGA CCTTACCAA TTATACAAA	TATAA ATTGC GAGAG CTGGG GCGAT
Cluster 3											
all 1608 m all 2167 m tll 2237 m tll 237 m tll 2398 m tll 2398 m tll 2398 m tll 2321 m tll 2321 m tll 2321 m tll 2031 m tll 2056 m tll 0256 m	CTATATCGGGT ACGTTGGTGAT CCTACCGCGCC CGATCTGCGGG ATCAGGCATCT TAAAATGGGGC TACTTTGGGGC TACGAACGCCT GAGGTCGGTAG GGGACCTCTAG	TTTATTCTTAT TCTGCGGGTAAT CGCCACCCCCA CGATCGCCTGCT CACGCCGCTGCCC AGGACGTGGTTT ACTGCGGAACGT CGGGAACGCCT CGATGCCTAAAGA	/TGAC//CAAAA /TGAC//CAATG /TGAC//CAATG /TGAC//CAATG /TGAC//CAATG /TGAC//CAATG /TGAC//CAATG /TGAC//CAATG	-GAATGGGAT -AAATGGGAT -AAATGGGAT -AAATGGGAT -AAATGGGAT -AAATGGGAT -AAATGGGAT -AAATGGGAT	FACAAGCCCCC FACAAGCTCCZ FACAAGCTCCZ FACAAGCCCCC FACAAGCCCCCC FACAAGCCCCCC FACAAGCCCCCC FACAAGCCCCCC FACAAGCCCCCCCCCC	GAGG T GCA-C T GCA-C T GCA-C T GCA-C T GCA-C T GCA-C T GCA-C T GCA-C T GCA-C	TAGAATCTCGGt TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT TTTAGGCCGGT	GAGTATGTCAR GAGTATGTCAR GAGTATGTCAR GAGTATGTCAR GAGTATGTCAR GAGTATGTCAR TGAGTATGTCAR GAGTATGTCAR GAGTATGTCAR GAGTATGTCAR	//CACGATCATT/ //AATCCAACAA/ //AATCCAATCA/ //ACCCCCTCT //AAGAACAAGT //AATGTCTCCAGA //TATCGCCAGAT A//CGTTTCCAG //AATTATGGTG/ //ATTATGCGCT	ACAACTGAA CCCTCTGGG 'ATCTTGCTG' 'ATCTTGCTG' 'AGACCGCAA GAACAAGGG 'TCCAACCTT 'TTGATCAGC AACTCACTA 'CAGTAATGC' 'TGTTACAGC	ACAGG TTAACC TTAAG ACCTC CTACT GATGC TGCGG TCTCCT TAACT CGTCC GCTTT
C Query Sbjct:	:46 ctgca       70068 ctgca	aatcaattta           aattaattte	aaattgcttt           caatttcttt	aatgcca        aatgccg	actatatc          actgtatc	gggttt      aagttt	ttattctta           tcattctta	at <u>tgac</u> aat          at <u>tgac</u> -at	cctagcgacc            cctaacgacc	c 117   c 69997	
D all7148 rn avaB0100 rr pcc7424	90 CTGAAGCATGAJ GCTGGTGAATT AAATAAATACC	100 1 <sup>-</sup> 	10 120 	130   . TAA-TGGAGC TAA-TGGAGC TAAATGGAGC	) 140 CAAGAGGACT CAAGAGGACT CAAGAGGACT CAAGAGGACT	CCGTTAC	150 AG AGCGGCGA AG AGCGGTGA AT GACGACT	1960 11 AGAGTGTCAA/, AGAGTGTCAA/, -GAGTGTCAA/,	970 1980   . (TTACTGTGCGGT (TTCCAAATTGAT; (CCGTTTTGTACA)	1990    IGCTTTTCAC ACCTATGCAG	) TATTTAA GAAGCTAG AATAAAAA
E		0 300 40		700 8	00 900	1000 11	00 1200 13	00 1400 15	00 1600 170	0 1800 1	Г 900 bp
avaB0101									$\rightarrow$		avaB0100
alr7149	_ <									$\rightarrow$	all7148
alr7009					=			$\Rightarrow$			all7008
F									V		
avaB0101 alr7149	ŧ			. <u> </u>	: =				= >		avaB0100
7424_0393	$\leftarrow$	-				: =				$\rightarrow$	7424_0394
alr7009	$\leftarrow$							$\longrightarrow$		_	all7008
G	1780	1790	1800	1810	1820	1830	1840	1850			
all3371 m	GACTGCAATCT	CTGTCtgaTGCC	TGAATCCCCCGT	TTTCAAAAC	GGGGGAG <mark>T</mark> GGC	TCAA//GO	GCTTACCTCTG	CCACGAA			
air7086 rn NpF6556 rn	GCCAGCAATCG GACTGCAATTG	CTGTTtgaGCCT CTGTTtgaAGCT	AGAA <mark>T</mark> CCCCCGA AGAA <mark>T</mark> CCCCCGA	ATTATATTCO ATTCAATTCO	GGGGGAG <mark>TAT</mark> G GGGGGAG <mark>TAT</mark> G	TCAA//AI	AGCCTCAATTCC CTATGTAATCAC	CTGAATTA CAGTACAA			

FIG. 4. Characteristics of IS891-related ISs. (/, presumptive left end of target site). (A) Border regions of copies of IS891. (B) Border regions of members of IS891-related clusters 2 (the top five) and 3. Lowercase tag and tga are the termination codons of the right-hand transposase ORF. (C) Result of a BLASTn search against the *A. variabilis* genome as target with, as query, the presumptive pretransposition region of IS(*alr2167*, *alr2168*) (see panel B, cluster 3), shows that tgac (underlined) is adjacent to the site of insertion of IS(*alr2167*, *alr2168*). (D) Border regions of members of IS891-related cluster 6, in which one cannot distinguish the L end (//, end of region of commonality) from a target site. (E) tBLASTx illustrates the greater similarity of IS(*all7148*, *alr7149*), IS(*avaB0100*, *avaB0101*), and a region from *Cyanothece* sp. strain PCC 7424, relative to IS(*all7008*, *alr7009*). In panels E and F, colors of bars correspond to alignment scores: red,  $\geq 200$ ; violet, 80 to 200; green, 50 to 80; blue, 40 to 50; and black, <40. ORFs are depicted as arrows with gray shafts from vertical lines to horizontal arrowheads. (F) The result of using IS(*all7008*, *alr7009*) as query supports the interpretation that IS(*all7386'*, etc.), inserted at the vertical arrowhead in panels E and F (see also Fig. 5, below), is within IS(*all7008*, *alr7009*). (G) The R border of IS(*all3371*, *alr3372*) in cluster 6 of the IS200-IS605 family shows an unusual example of an IS891-related IS in which the R end, near the (lowercase) termination codon tga of *alr3372*, may terminate in a variation of TGTCAA.



FIG. 5. Stacked ISs. (A) A linear portrayal of a region of plasmid  $\alpha$  in the vicinity of its arbitrary origin, a unique SalI site within *all7001*. Except for *alr7007*, ORFs *asr7385* through *alr7009* are annotated as encoding transposases. The ends of the ISs are indicated by individualized brackets. (B) A view of the same region illustrating that IS(*all7008*, *all7009*) has, at the position of the vertical arrowhead in Fig. 4E and F, been interrupted by insertion of IS(*alr7386'*, *alr7003'*, *asr7006*, *alr7007*), itself interrupted by IS(*all7002*, *all7001*) and by IS(*all7003*, *all7004*). Insertion of the latter two ISs added stop codons that subdivided IS(*alr7386'*, etc.) into its several "ORFs," *"alr7386*," *"alr7003*," and *"asr7006*." The horizontal arrowheads represent IRs at the ends of the ISs, and the series of letters next to some of these arrowheads represent DRs flanking the IRs.

transposases (RAYTs) associated with repetitive extragenic palindromes in inverted orientation are found in diverse bacteria, show a relationship to IS200 (22), and match well with Alr1015 (Fig. 6A).

*A. variabilis* has no substantial fragment of IS(*alr4734*) in its genome. When the region of IS(*alr4734*), computationally freed of the region between its inverted repeats, is used as a query with *A. variabilis*, a long region of homology is found to both sides of the IS, approaching to within 8 bp of the IR on the left and to within 68 bp on the right. These data suggest that one need expect no sequence other than the *alr4734*-containing region bracketed by the IRs to be required for transposition (data not shown). Some of the ends of related ISs, e.g., those of IS(*alr1015*), are partially palindromic (<u>GGT GGGCAATGCCCACC</u>), with the palindrome (underlined) reaching to within 4 bp of the end of the IR, whereas others (e.g., those of IS[*alr4734*]) are not palindromic (Fig. 6B). It is unclear whether parts of the IRs are target rather than parts of the ISs.

**IS630 family.** IS895 and its relatives (1) are part of the IS630 family, whose members normally have one ORF (ISfinder). However, seven of the eight *Anabaena* sp. IS895s bear pairs of nonidentical transposases. Indel (insertion-deletion) mutations of adenylic acid residues in a poly(A) stretch of nucleotides, positions 594 to 603 in Fig. 7A, lead to changes in the site of translational termination. Just such a sequence in IS895 is thought to incur translational frameshifts (32). A similar indel mutation is seen in Fig. 7B for positions 386 to 393 of IS891 cluster 2.

IS630 family members normally have terminal IRs that are flanked by 2-bp DRs. As we found for IS(*all7564*, *all7563*) in cluster 5 (Fig. 8C and D), their target site is "often CTAG with duplication of the TA" (ISfinder). More generally, we found repeatedly in the IS630-related ISs that we studied, but not elsewhere, that the presumptive ends of ISs terminated in short palindromes, for example, TATA, shown as part of an IR in Fig. 8A1. However, the same tetranucleotide could be considered a DR (Fig. 8A5) or could be subdivided into TAs outside the IS and TAs inside the IS, as in Fig. 8A3. If one were to remove such an IS and one copy of its DRs computationally (Fig. 8B), the resulting sequences would retain different numbers of As and Ts, and the reading frame of the predicted protein (if the IS was within a protein-encoding gene) would change (right side of Fig. 8B). Therefore (see Materials and Methods), BLAST analysis was used to distinguish whether those palindromes are part of the IRs, actually DRs, or part of both. For IS(alr1726, alr1727) in cluster 1, the only interpretation that matches Fig. 8E and has a DR duplicates TA. With that small amount of evidence added to the data and conclusion presented by ISfinder, we tentatively assign TA DRs for all five clusters of IS630 members of Anabaena sp. (Fig. 8F and Table 1).

The IS892 cluster of ISs. IS892 (7, 9) is present in ISfinder within an assemblage of "unclassified IS elements" (17). Whereas copies of IS1594 and IS895 are present only in the chromosome of *Anabaena* sp., copies of IS892 are present only in plasmids  $\alpha$  and  $\delta$  of that strain. The latter localizations and the ability of plasmid  $\alpha$  to move from one strain to another (21) suggest strongly that IS892 reached *Anabaena* sp. within a plasmid. Ten copies of IS892 bearing 18 *Anabaena* sp. ORFs show similarity to each other. Although nearly identical in nucleotide sequences, these copies have one, two, or three ORFs so labeled within a single copy (Fig. 7C and Table 1). Like the copies of IS895, those of IS892 cluster 1 have series of poly(A) residues that differ among copies; differences between those series result in variations in the sites of termination and

### A Alignment of Alr1015 and homologs

Alr1015   Imperative construction of the intervent of the interv		10	0 2	0	30	40	50	60	7	0	80	90	
Alr1015   MEPEYRRAYLPGGTFFTLTVTERYPIFENTENTSHLRASALAKVRSEMPFELGAVVLPDHHFTWTLPPEDDRNYSQRIGELKVLFTRSJDSKTLIPK     Cxb07424   MENYRRLYIPGGTYFTLVTYRRYPIFENSENTSALRQAVAVKKQMPFEILGAVLLPDHHFTWTLPEKDRNYSKRIGEIKUFTHALRQANNLPE     CceOffa   MENYRRYIPGGTYFTLVTYQRKPYFDNSDNIKLRQAVAVKKQMPFEILGAVLLPDHHFTWTLPEKDRNYSKRIGEIKUFTKSFRGNNALPK     Alr4734   MSNYRRYIPGGTYFTUVTYQRKPYFDNSDNIKLRQAVATVKKEMPFNILGAVLLPDHHFTWTLPPEDDNYSKRIGEIKKKLFTKSFRGNNALPK     Alr4734   -MDYRRAKIEGGTFFTUVTYQRKPYFDNSDNIKLRQAVATVKKEMPFNILGAVLLPDHHFTWJLPPEDSNFSMRRILKNYFTRICSKYKQETI     CceOff   MSNYRRLYFKGSTFTUVTYQRKPYFDNSDYTKLRIAVSQVKKEMPFNILGAVLLPDHHFTWSLPRKDNYSQLISRKKULFTKSFGRQLRSVN     Mic.chth   MSNYRRYIPGGTYFTUVTQRKPYFDNSDYTKLRIAVSQVKKEMPFNILGAVLLPDHHFTWSLPRKDNYSQLISRKKULFTKSFGRQLRSVN     Mic.chth   MSNYRRYIPGRCYFFTUVTQRKPYFDNSDNINKLRIAVATVKKEMPFNILGAVILPDHHFTWTLPFCDDDFSSRKRIKKQFT													••
Cyth7424 MPNYRLYTPGGTYFFTUTYTROPHYSSENDALDQAVAWYKKWPFTLGAVILPDHIHFUMTPERDFNYSKIGIKVLFTHALRIQNNLPE CceOrf MSNYRRJYTPGGTYFFTUTYTROPHYFANSDNIKLRQAVAWYKKWPFTLGAVILPDHIHFUMTPERDFNYSKIGIKKVLFTHALRIQNNLPE MSNYRRJYTPGGTYFFTUTYTROPHYFANSDNIKLRQAVAWYKKWPFTLGAVILPDHIHFUMTPERDPNYSKIGIKKNESTKIGREKNNTLPD Alr4734MDYRAAKIEGGTFFTUTYTROPHYFANSDNIKLRQAFRQVIAQVPFTVDATULPHHHCUMTPPCDSNESNRWALIKNYFTRGSTKYGET CceOrf MSNYRRJYTRGGTYFFTUTYTROPICKNSQUISQKEEKPFDIGAVILPDHIHFUMTPERDSNYSQUISRKKULFTRSGRACLSKYGET CceOrf MSNYRRJYTRGGTWFFTVDIKNRRSQLITTQYQMLRHAI-KVRRPFFUDAIVILPDHIHFUSTPHDDYSQLISRKKULFTRSGRACLSKYGET 	Alr1015	MPEYRRAYLP	GGTFFLTLV	<b>FYERYPIF</b> S	SNIENISHI	LRSALAKVR	SEMPFEIE	AVVLPDHI	HFLWTLPT	DKNYSQRI	GRLKVLFT	SIDSKTII	PK
CceOdfA   MSNYRRLYTPGGTYFFTUTYKRPYFANSDNIIKLRJAVATYKKEMPFNILGAVILPDELLEFLWTLPPEDNNYSKRIGLIKASFTKLFRGNNILPK     Alr473   -MSNYRRYTPGGTYFFTUTYGRVPFDNSDNINKLRIAVATYKKEMPFNILGAVILPDELLEFLWTLPPEDNNYSKRIGRMKLFTKSFRGNNILPK     Alr474   -MSNYRRLYFKGSTFFITUTYGRVPFENDDNINKLRIAVATYKKEMPFNILGAVILPDELHFLWTLPPEDNNYSKRIGRMKLFTKSFRGNNALPK     Mic.chth   MSNYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSGYKKEMPFDILGAVILPDHLHFLWTLPPEDDNYSKRIGRMKLFTKSFRGNNALPK     Ecol   MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSGYKKEMPFDILGAVILPDHLHFLWTLPPEDDNYSKRIGRMKLFTKSFRGNNALPK     Ecol   MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSGYKKEMPFDILGAVILPDHLHFLWTLPPEDDNYSKRIGRMKLFTKSFRGNNALPK     MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSGYKKEMPFDILGAVILPDHLHFLWTLPPEDDNYSKRIGRMKLFTKSFRGNNALPK   -     MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSKEMPFDILGAVILPDHLHFLWTLPPEDDNYSKRIGRMKLFTKSFRGNNALPK   -     MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSKEMPFDILGAVILPHDHLFWTEGVSKEMVOYSKRIGTKMLFTKSFRGNNALPK   -     MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSKEMPFDILGAVILPHOLIWTSKRIGTKMKLFTKSFRGNNALPK   -     MSYRRFYTEGGTYFFTUTYGRVPFENDDVITKLRIAVSKEMPFDILGAVILPHTMTLFTKSFRGNNALPK   -     Alr1015   NLSNSRKHRESNWQRRFWEHTLQEESELTYLNYTHYFYKHKLVSCPHLWVSSPHTVKKKSFDWACICSCGCFKKTUDENDELKNYCE     CceOffA   TSSRSKHRESNWQRRFWEHTLQEESELTYLNYTHYFYKHKLVSSPHWVKKGYDSSNWCCVQGCHLDKINDELKNYCE     CceOffA   -	Cyth7424	MPNYRRLYIP	GGTFFFTLV	<b>TYHR</b> QPH <b>F</b> S	SSS <mark>ENVAKI</mark>	LRQAVAMVK	KQMPFEIL	AVILPDHI	HFLWTLPE	<b>KDFNYSKRI</b>	GRIKVLFT	ALRNQNNI	PE
CceOrdB   MSNYRRFYLFGGTYFFTUTYDKRFYLFDNSDNINKLRIAVATVKKEMPFNILGAVILPDBLHFLWTLPPDDNSYSRITGRMKMLFTKSFRGNNALFK     Alr4734   -MYRRAKIEGGTFFTVVTHNRRFLCEPENIRLLRQAFRQVIAQYPFTVDATVILPNHHELWTLPPDDNSYSRITGRMKMLFTKSFRGNNALFK     CceOrdB   MSNYRRFYLFGGTYFFTVTYDRNSDNINKLRIAVATVKKEMPFNILGAVILPDHLHFLWTLPPDDNSYSRITGRMKMLFTKSFRGNNALFK     Mic.chth   MSNYRRFYLFGGTYFFTVTYDRVFFTVNINRRSQLLTTQYQMLRHAII-KVKEMPFNILGAVILPDHLHFLWTLPPDDNSYSRITGRMKMLFTKSFRGNNALFK     Ecol   MSYRRFYLFGGTYFFTVNINRRSQLLTTQYQMLRHAII-KVKEMPFNILGAVILPDHLHFLWTLPPDDNSYSRITGRMKMLFTKSFRGNNALFK     Alr1015   100   120   130   140   150   160   170   180   190     Alr1015   NLSNSRKHRESNWQRRFWENILDEADFERLMYLHYNPKHGLVSCHLWYSFFKYRGIYGSNWCSCGRFUDIPDFNLKKYGE   A   100   100   120   130   140   150   160   170   180   190     Alr1015   NLSNSRKHRESNWQRRFWENILDEADFERLMYLHYPKHGLVSCHLWYSFFWRGIYGSNWCSCGRFUDIPDFNLKKYGE   100   110   120   120   160   170   180   190     CceOrdF   TVSTSRQKHRESNWQRRFWENTLDQEEELTYLNYHYPKHGUVSCPHWPSSFQTWVGKKAYSFDWACICHQCVLNDKINFNFDHLKYVGE   SSTRGVRGSNWCCCQCOLDKINFNFDHLKYVGE   100   100   100   100   100   100   100   100   100   100   100   100   <	CceOrfA	MSNYRRLYIP	GGTYFFTLV	TYKRTPYF7	ANSDNIIKI	<b>IRQAVATVK</b>	YEMPFNIL(	AVILPDHL	HFLWTLPP	DNNYSKRI	GLIKASFT	LFRGNNTL	PD
Alr334   -MDYRRAFIEGGTFFFTVVTUNRREFLCEPENIRLLRQAFRQUTAQVYFTUDATULIPNIHHCUWTLPPGDSNFSNRWRLIKNYFTHCGIKYQETI     CceOrfC   MSNYRRFYIPGGTYFFTUTYQRVPLFKNDYITKLRIAVSQVKKEKPFDIIGAVILPDHHFIWSLPKHDSNYSQLISRMKVLFTRSIGRQLRSVN     Mic.chth   MSYRRFYIPGGTYFFTUTYQRVPLFKNDYITKLRIAVSQVKKEKPFDIIGAVILPDHHFIWTLPPHDQNYSKRIGRMKHLFTKSFRGNALPK     Ecol   MSYRRFYIPGGTYFFTUTYQRVPLFKNDYITKLRIAVSQVKKEKPFDIIGAVILPDHHFIWTLPPHDQNYSKRIGRMKHLFTKSFRGNALPK     MSYRRFYIPGGTYFFTUTYQRVPLFKNDYITKLRIAVSQVKKEKPFDIIGAVILPDHHFIWTLPPHDQNYSKRIGRMKHLFTKSFRGNALPK     Ecol   MSYRRFYIPGGTYFFTUTYQRVPLFNDDNINKLRIAVATVKKEMPFTIIGAVILPDHHFIWTLPPHDQNYSKRIGRMKHLFTKSFRGNALPK     MSYRRFYIPGGTYFFTUTYQRVPYPDSDNINKLRIAVATVKKEMPFTIIGAVILPHHHFUKTPFDDDNYSKRIGRMKHLFTKSFRGNALPK     Ecol   100   110   120   130   140   150   160   170   180   190   1	CceOrfB	MSNYRRFYIP	GGTYFFTIV	TYQRKPYFI	DNSDNINKI	<b>GRIAVATVK</b>	KEMPFNII(	AVILPDHL	HFLWTLPPN	<b>DQNYSKRI</b>	GRMKMLFTF	SFRGNNAL	PK
CceOrdC   MSNYRRLYKGSTIFLTUTUTUGRVELFKNSDYITKLRIAVSQVKKENFPILGAVILPDHLHFLWSLPKHDSNYSQLISKMKVLFTKSIGRQLKSVN     Mic.chth   MSNYRRLYIGGTWFFTVLRNRRSQLLTUTUYQRVELFKNSDYITKLRIAVATVKKEMPFNIGAVILPDHLHFLWTLPPNDQNYSKRIGRMKMLFKSFRGNNALPK     Ecol   MSNYRRLYKGSTWFTVNLRNRRSQLLTUTUYQRVELKNAII-KVKRDRPFEINAWVLPEHHCIWTLPPEGDDPSSRWEIKKQFT	Alr4734	-MDYRRAKIE	GGTFFFTVV	HNRREFLC	CEPENIRLI	RQAFRQVI	AQYPFTVD2	IVILPNHI	ICLWTLPP(	DSNFSNRW	RLIKNYFTF	HCSIKYQE	TI
Mic.chth MSNYRRFY1FGGTYFFTIVTYQRRFYEDNSDNINKLRTAVATVKKEMPFNIIGAVILPDHLHFLMFLPPDDNYSKRIGAMKNLFYKSFGNNALPK Ecol MSNYRRFY1FGGTYFFTIVTYQRRFYEDNSDNINKLRTAVATVKKEMPFNIIGAVILPDHLHFLMFLPPDDDNYSKRIGAMKNLFYKSFGNNALPK MSYRRY1KGGTWFFTVNLRNRSQLTTQYMLRHAII-KVKRDRPFFINAWVULPEHMHCIWTLPBGDDFSSRWFIKQFT	CceOrfC	MSNYRRLYFK	GSTIFLTIV	<b>FYQRVPLF</b>	KNS <mark>DYITKI</mark>	LRIAVSQVK	KEKPFDIL(	AVILPDHL	HFIWSLPK	HDSNYSQL]	SRMKVLFT	RSI <mark>G</mark> RQLRS	VN
Ecol   MSEYRRYYIKGGTWFFTVNLRNRRSQLLTTQYQMLRHAII-KVKRDRPFEINAWVULPEHMCIWTLPEGDDDFSSRWREIKKQFT	Mic.chth	MSNYRRFYIP	GGTYFFTIV	ryqrkpyf <mark>i</mark>	ONSDNINKI	LRIAVATVK	KEMPFNII(	AVILPDHL	HFLWTLPPI	DQNYSKRI	GRMKMLFT	SFRGNNAL	PK
A I A IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Ecol	MSEYRRYYIK	GGTWFFTVNI	LRNRRSQLI	LTTQYQMLF	RHAII-KVK	RDRPFEIN	WVVLPEHM	HCIWTLPE	DDDFSSR	REIKKQFT-		
Alr1015 Cyth7424 CceOrfA Cc		<b>A</b>	I.	<b>A</b>				▲* '	* 🔺 📔		<b>A</b>		
Alr1015   1100   120   130   140   150   160   170   180   190     Alr1015   NLSNSRKHRESNVWQRFWEHTLQDEADFEKHLNYIHYNPVKHGLVSCPHLWDYSSFHKFVRGTYCSNWYCSCSGERIQIPDFDKNLEKIGE     Cyth7424   NISNSRKHRESNVWQRFWEHTLQDESELETYLNYIHYNPVKHGLVSCPHLWDYSSFHKFVRGTYCSNWYCSCSGERIQIPDFDKNLEKIGE     CceOrfB   TVSISRQKHRESNVWQRFWEHTLQDESELETYLNYIHYNPVKHGLVSCPHLWLYSSFHKFVQKDSYSNWGCICQGQFKNIDNFDLKNYVGE     Alr4734   -STSRQSNKGELAVWQRFWEHTIQDESELETYLNYIHYNPVKHGLVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     CceOrfC   NIPESRKKRESNVWQRFWEHTIRNEQDLKTHLDYIHYNPVKHGVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     Mic.chth   TVSTSRQKHRESNVWQRFWEHTIRNEQDLKTHLDYIHYNPVKHGVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWQPRFWEHTIRNEQDLKTHLDYIHYNPVKHGVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     All A III   *A     All A IIII   *A     SACCTTGGCCAATGTCAA/GGTGGGCAATGCCCACCAAAAAAATATTATG     CceOrfA   GACCTTGGCCAATGCAATGCCCACCACAAAAAAATTTAATGAAAAAAATATATAT		1.0.0						<b>A</b> 4	<b>A</b>				
Akr1015   NLSNSRKHRESNVQQRFWEHTIQDESLETVLNYIHYNPVKHGUSCPHLWDYSSFKFVRGYCSCSGERIQIPDFDKNLEKIGE     Cyth7424   NLSNSRKHRESNVQQRFWEHTLQDESLETVLNYIHYNPVKHGUSCPHLWPYSSrQTWCKKAYSPDWACTCHRQKVLIPNFDLIDKVGE     CceOrfA   TVSTSRQKHRESNVWQRFWEHTLQDESLETVLNYIHYNPVKHGUSCPHLWLYSSFHTWVQKDSYSSNWGCYQQQNLDKIPNFDLILDKVGE     Alr4734   -STSRQSNKGELAVWQRFWEHTLQDESLETVLNYIHYNPVKHGUSCPHLWLYSSFHTWVKKEGYSSNWGCYQQQNLDKIPNFDNLNYVGE     CceOrfC   NIPSSRKHRESNVWQRFWEHTIQDESLETVLNYIHYNPVKHGUSCPHLWLYSSFHTWVKKEGYSSNWGCYQQNLDKIPNFDNLNYVGE     Mic.chth   TVSTSRQSNKGELAVWQRFWEHTIQDESLETVLNYIHYNPVKHGUSCPHLWLYSSFHTWVKKEGYSSNWGCYQQQNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWQRFWEHTIQDESLETVLNYIHYNPVKHGUSCPHLWLYSSFHTWVKKEGYSSNWGCYQQDULKIPNFDNLKNYVGE     Ecol  HACGLKNIWQPRFWEHTIQDESLETVLNYIHYNPVKHGUSCPHLWLYSSFHTWVKKEGYSSNWGCYQQQLDKIPNFNUKNYVGE     alr1015 m   GACCTTCGCCAAATAATCCAA/GETGGGCAATGCCCACCACAAAAATATTAG     cceOrfM   A   I   *A     ATCGTAAGGACAATGAAAGATAAT/GTAGGGTGGGGCAATGCCCACCCACACAAAAATATTAG   TAATGGTGGGCAATGCCCACCCTAC//TCAATATCAAGAATAAATGAATT     cceOrfM   ATCGTAAGGACGAATGACGAATGCGCCACCCCCCCCCACAAAAATATTAG   TAATGGTGGGGCAATGCCCACCCTAC//TCAATATCAAGAAATAATTAAGATTAATGCAATGGTGGGGCAATGCCCACCCTAC//TCAATATCAAGAATAAATGCCAACGCCACCCTAC//TAACAATGCAATG		100	110	120	130	14	U .	.50	160	T / 0	180	190	
AITOTS NLSNSRKHRESNVWQRFWEINTRDEADFERHIN/ HYNPVKHGLVSCPHLWDYSSFHRFVRRGT/SCSGERGUPPDFDRILERTE Cyth7424 NISNSRKHRESNVWQRFWEITLQDESELETYLNYTHYNPVKHGLVSCPHLWDYSSFQWVGKRAYSFDWACICHRQKVLTPNFDLITDKVGE CceOrfA TVSTSRQKHRESNVWQRFWEITLQDESELETYLNYTHYNPVKHGLVSCPHLWLYSSFHRVVRKGYSSNWGCUCQGNLDKTPNFDNLKNYVGE Alr4734 - STSRQSNKGELAVWQRFWEITLQDESELETYLNYTHYNPVKHGLVSCPHLWLYSSFHRVVRKEGYSSNWGCUCQGNLDKTPNFDNLKNYVGE CceOrfC NIPESRKHRESNVWQRFWEITLQDESELETYLNYTHYNPVKHGUVSCPHLWLYSSFHRVVKKEGYSSNWGCUCQGNLDKTPNFDNLKNYVGE Mic.chth TVSTSRQKHRESNVWQRFWEITTINEQDLKTHLDYTHYNPVKHGUVSCPHLWLYSSFHRVKKEGYSSNWGCUCQGNLDKTPNFDNLKNYVGE EcolHACGLKNTWQPRFWEITTINEQDLKTHLDYTHYNPVKHGVVSQVSDW	4-4045				.				•• ••••				
CceOrfA TVSISRKHRESNVWQRRFWEHTLQDESELETYLNYHYNPVKHGLVSCPHLWPYSSFDTWVGKKAYSFDWACICHRGKVLIPNFDLIIDK/GE CceOrfB TVSISRQKHRESNVWQRRFWEHTLQDESELETYLNYHYNPVKHGLVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQQNLDKIPNFDNLKNYVGE Alr4734 -STSRQSNKGELAVWQRRFWEHTIDDESELETYLNYHYNPVKHGLVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQQNLDKIPNFDNLKNYVGE CceOrfC NIPESRKHRESNVWQRRFWEHTIRNEQDLKTHLDYHYNPVKHGVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQQNLDKIPNFDNLKNYVGE Mic.chth TVSTSRQKHRESNVWQRRFWEHTIRNEQDLKTHLDYHYNPVKHGVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE EcolHACGLKNIWOPRFWEHAIRNKDYRHVDYIYINPVKHGVVQVSDWPSTFHRDVARGLYPI-DWAGD AA A III X IIII X A III X A IIII X A III X A X A	AIFTUTS	NLSNSRRKH	IRESNVWQRR	WEHNIRDE	SADFERHL	NYTHYNPVK	HGLVSCPH.	WDYSSFHK.	EVRRGIYC:	SNWYCSCSC	ERIQIPDET	KNLEKIGE	
CCeOOTA   TVSTSRQKHRESDVMQRRFWEHTLQDESELETYLNYLHYNPVKHGLVACHLWLYSSFHTWVQRDSYSSNMGCICQGQFKNIDHDELKDVGE     CCeOOTA   TVSTSRQKHRESDVMQRRFWEHTLQDESELETYLNYLHYNPVKHGLVSCPHLWLYSSFHTWVQRSGSNMGCICQGQFKNIDHDELKDVGE     AIr4734   -STSRQSNKGELAVWQRRFWEHTLQDESELETYLNYLHYNPVKHGVVSCPHLWLYSSFHTWVKKEGYSSNMGCCCQGNLDKIPNFDNLKNYVGE     CceOrfC   NIPESRKHRESNVWQRRFWEHTIRNEQDLKTHLDYIHYNPVKHGVSCPHLWLSSFHTWVKKEGYSSNWGCCCQGNLDKIPNFDNLKNYVGE     Mic.chth   TVSTSRQKHRESNVWQRRFWEHTIQDESELETYLNYIHYNPVKHGLVSCPHLWLSSFHTWVKKEGYSSNWGCCCQGNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWOPRFWEHAIRNTKDYRHHVDYIYINPVKHGVKQVSDWPFSTFHRDVARGLYPI-DWAGD     AA   A     A   A     AA   A <t< th=""><th>Cyth/424</th><th>NISNSRKKH</th><th>IRESNVWORR</th><th>WEHTLOOF</th><th>SEELELYLI</th><th>NYTHYNPVK</th><th>HGLVSCPH.</th><th>WPYSSEQT</th><th>WVGKKAYS</th><th>DWACICHE</th><th>QKVLIPNFI</th><th>DLIIDKVGE</th><th></th></t<>	Cyth/424	NISNSRKKH	IRESNVWORR	WEHTLOOF	SEELELYLI	NYTHYNPVK	HGLVSCPH.	WPYSSEQT	WVGKKAYS	DWACICHE	QKVLIPNFI	DLIIDKVGE	
CceOrds   TVSTSRQHRESNV@QRFWEHTIQDESLETVLNTHYPVRHGLVSCPHLWLYSSFHTWVRREGYSSNWGCVCQGNLDKIPHPDNLRNVGE     Alr4734   -STSRQSNKGELAVWQRRFWEHTIRNEQDLKTHLDYIHYPVRHGLVSCPHWEYSSFHTWVRREGYSSNWGCVCQGNLDKIPHPDNLRNVGE     CceOrfC   NIPESRRKHRESNVWQRRFWEHTIRNEQDLKTHLDYIHYPVRHGLVSCPHWEYSSFHQWVRRGFYPLEWGCCCHGRSTKTPDFSNIINQVGE     Mic.chth   TVSTSRQKHRESNVWQRRFWEHTIQDESLETYLNYIHYPVRHGLVSCPHWEYSSFHTWVRREGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWQPRFWEHTIQDESLETYLNYIHYPVRHGUVSCPHWEYSSFHTWVRREGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWQPRFWEHAIRNTKDYRHHVDYIYINPVRHGUVSCPHWEYSSFHTWVRREGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     B Border regions  HACGLKNIWQPRFWEHAIRNTKDYRHHVDYIYINPVRHGUVRQVSDWPFSTFHRDVARGLYPI-DWAGD     alr11   *A	CceOrfA	TVSISRQKH	IRESDVWQRR	WEHTLODE	ESELEIYL	NYIHYNPVK	HGLVACPH.	WLYSSFHT	WVQKDSYS	SNWGCICQC	QFKNIDNFI	DELKDYVGE	
Alf4734   -STSRQSNKGELAVWQRFWEHQIKNEVDFTHHFYIHYNPVKHGVKAPKDWQYSSFILYQQREFYNIDWGTGEDIKFSQDIGNE     CceOrfC   NIPESRKHRESNVWQRFWEHTIRNEQDLKTHLDYIHYNPVKHGVSCPHWEYSSFHQWVKRGFYPLEWGCCCHGKSTKTPDFSNIINQVGE     Mic.chth   TVSTSRQKHRESNVWQRFWEHTIRNEQDLKTHLDYIHYNPVKHGVSCPHWEYSSFHQWVKRGFYPLEWGCCCQGNLDKIPNPDNLKNYVGE     Ecol  HACGLKNIWQPRFWEHTIQDESELETYLNYIHYNPVKHGVVSQVSDWPFSTFHRDVARGLYPI-DWAGD     AAA   AIII   *A   III     AAA   AIIII   *A   IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		TVSTSRQKH	IRESNVWORR	WEHTIQDE	SELETYLI	NYTHYNPVK	HGLVSCPH.	WLYSSFHT	WVKKEGYS	SNWGCVCQC	NLDKIPNFI	NLKNYVGE	
Cccord   NTPESRRARESNVWQRFWEHTTRNEQDLKTHLVTHYNPVKHRLVSCPHHWEYSSPHQwVRRGFYPLEWGCCCHGKSTKTPDPSNIINQVGE     Mic.chth   TVSTSRQKHRESNVWQRFWEHTTQDESELETYLNYIHYNPVKHGLVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWQPRFWEHAIRNTKDYRHHVDYIYINPVKHGVSCPHLWLYSSFHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     B Border regions   60   70   80   90   100   110   120   990   1000   101   1020   1030   1040     alr1015 m   GACCTTCGCCAAATAATCCAA//TTAAGGATGGGCAATGCCCACCAAAAAATATTAG   TAATGGTGGGCAATGCCCACCCAAAA   TAATGGTGGGCAATGCCCACCCAAAA   TAATAGGTGGGCAATGCCCACCCAAAA   AAAAAGGTGGGCAATGCCCACCCACCAAAA   AAAAAGGTGGGCAATGCCCACCCAAAA   AAAAAGGTGGGCAATGCCCACCCAAAA   AAAAAGGTGGGCCAATGCCCACCCACAAAAAAATAATTAAAATAAAATAAAATAAAATTAAATTAAATTAAATTAAA   AAAAGGTGGGCAATGCCCACCCACCAAAAAAATAATTAAATTAAATTAAATTAAATTAAATTAAAA	Alr4734	-STSRQSNK	GELAVWQRR	WEHQIKNE	EVDETHHET	NYIHYNPVK	HGYVKAPK	WQYSSFIL	YVQRERYN.	LDWGTO	EDIKES	QDIGNE	
Mic.chth   TVSTSRQKHRESNVMQRFWEHTTQDESELETYLINTHYNPVKHGLVSCPHLWLYSSPHTWVKKEGYSSNWGCVCQGNLDKIPNFDNLKNYVGE     Ecol  HACGLKNIWOPRFWEHAIRNTKDYRHVDYIYINPVKHGVXQVSDWPFSTFHRDVARGLYPI-DWAGD     AA   A   III   *A   IIII   IIII   IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	CceOrfC	NIPESRRKH	RESNVWORR	WEHTIRNE	EQDLKTHLI	DYIHYNPVK	HRLVSCPH	IWEYSSFHQ	WVKRGFYPI	LEWGCCCHO	KSTKTPDFS	NIINQVGE	
Ecol  HACGLKNIMOPRFWEHAIRNTKDYRHHVDYIYINPVKHGWVKQVSDWPFSTFHRDVARGLYPI-DWAGD     AA   A   I   *A   III   IIII   IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Mic.cnth	TVSTSRQKH	RESNVWQRRI	WEHTIQDE	ESELETYLN	<b>YIHYNPVK</b>	HGLVSCPH	WLYSSFHT	WVKKEGYS	SNWGCVCQG	NLDKIPNFI	NLKNYVGE	
B Border regions 60 70 80 90 100 110 120 990 1000 1010 1020 1030 1040 1	Ecol	HAC	GLKNIWOPRI	WEHAIRNI	TKDYRHHVI	YIYINPVK	HGWVKQVSI	W		PFSTE	HRDVARGLY	PI-DWAGD	
B Border regions <sup>60</sup> 70 80 90 100 110 120 990 1000 1010 1020 1030 1040 <sup>60</sup> Crostractartcartcartcartcartcartcartcartcart						^ <b>A</b>		I					
B Border regions $\begin{bmatrix} 60 & 70 & 80 & 90 & 100 & 110 & 120 & 990 & 1000 & 1010 & 1020 & 1030 & 1040 \\ \hline \\ alr1015 m \\ cce7424 m \\ cce07fA m \\ alr4734 m \\ alr4734 m \\ credf a ccatacatacatacatatatatatatatatatatata$						<b>A</b>							
B Border regions     60   70   80   90   100   110   120   990   1000   1010   1020   1030   1040     alr1015 m   Gacctrccacaaraatrcaatccaartaatrcaatctaa//gtagegegegegegegegegegegegegegegegegegeg													
alr1015 m   GACCTTCGCCAAATAATTCAACCATTCAAGGGTGGGCAATGCCCCACCAAAAAATATTAAG   TAATGGTGGGCAATGCCCACCCTAC//TTAAGAAATTTAAGAAATATAAAAATAAAAT	<b>B</b> Borde	r regions											
alr134 m GGGTTGACCATAGCATTGAGCATGGCGAAAC/GTAGGGTGGGCAAATGCCGCACCAAGAAAC/GTAGGGTGGGCAAATGCCGACGCAGGAGAAATAATCGAACGAA		70	0.0	0.0	100	110	120	000	1000	1010	1020	1020	1040
alr1015 m GACCTTCGCCAAATAATTCAATCCAA/GTAGGGTGGGCAATGCCCACCAAAAATATTATG cce7424 m CTGTTACTGTAACTAAAAAGTAAT/GTAGGGTGGGCACTGCCCACCACAAAAATATTATG cceOrfA m ATCGTAGTTAAAGAGAAGAGAATGAGGAATGAGGAATGACGATTGCCCACCATAGCAAGAAAAAATGATTG cceOrfB m AATCGTAGCTAAGGAGGAGGCACTGCCCACCCTTAACCAATGGAGGAGAAATAA alr4734 m GAAGCGTAACGGTTAAAATATAGT//GTAGGGTGGGCAATGCGCGCGGGAGAGAAATAA alr4734 m GAAGCGTAACGATTAAAAAATATAGT//GTAGGGTGGGCAATGCGCGCGGGAGAGAAATAA alr4734 m GAAGCGTAACGATTAAAAAATATAGT//GTAGGGTGGGCAATGCGCGCGGGAGAGAAATAA alr4734 m GAAGCGTAAACGGTTAAAATATAGT//GTAGGGTGGGCAATGCGCGCGGGAGAGAAATAA alr4734 m GAAGCGTAACGATTAAAAAAAAAAAAAAAAC//GTAGGGGGGAAATACTGGGAGAGAAATAAA	00  .	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,				 	120						
cce7424 m   CTGTTTACTGTTACTGTAACTAAAAAGTAAT//GTAGGGTGGGCACTGCCCACCACAAA   AAAAAGGTGGGCAGTGCCCACCTAC//GGTTGTAATTAAGAAAAAAATAGATTT     cceOrfA m   ATCGTAGTTAAAGAACGAATGAGGAATT//GTAGGGTGGGCAATGCCCACCATAACCATTGCT   ATAATGGTGGGCAATGCCCACCCTAC//TTAACTAACGAACTGAGG     cceOrfB m   AATCCTAACTAGAAGCACTTAAATAA//GTAGGGTGGGCAATGCCCACCCTTAACCAATGCA   AGTAAGGTGGGCCACCCTAC//TTAACTAACTAACTAACTGAGGG     alr4734 m   GAAGGCGTAACGGTTTAAATAAAGT//GTAGGGTGGGCAATGCCGCGGGAGAGAAATAA   CTGCACCCAACCCACCCCAC//TAACTAACTGCCACCCGTCA     cceOrfC m   GGTTTCAACAATACCATTGCGCGGCGAAATACTGGTAGCGCGTGAGCAAATACTGTTAACTACCGACCG	alr1015 m G	ACCTTCGCCAAATA	ATTCAATCTCA	//GTAGGGTG	GGCAATGCC	CACCAAAAAT	ATTATG 1	AATTGGTGGG	CAATGCCCA	CCCTAC//TC	AATATTCAAG	ATTTAATAA	ACTAA
cceOrfA m ATCGTAGGTAAAGACGAATGAGGATT//GTAGGGTGGGCATTGCCCACCTTAACCAATGCT ATATATGGTGGGCAATGCCCACCTAC//TTAACTTAAAATGTAATCCAACTGGTG   cceOrfB m AATCCTAACATGAGCACTTAAATAAA/ATAGGTGGCCAACGCCACCCTTACCAACTGCAAGAATAA AGTAAGGGGGCAATGCCCACCCTAC//TTAACTAACTACTGCAATAAATCC   alr4734 m GAAGGCGTAACGGTTAAATATAGT//GTAGGGTGGGCAATGCCGCGGGAGAGAAATTAA TCTACACTCGGTCCCACCCTAC//TAACTAACTGCGACGCCGCGCGAGAGAAATTAA   cceOrfC m GGTTTCAACATACCAATGCAAATAAAGT//GTAGGGTGGGCAAATACTGTAACCGTTGACCAACCTAC//TAACTAACTGCGACGCCCTAC//TAACTAACTGCCGACCGCCAATGCCCCCCCAACCCTAC//TATTAACTTCTAACTACCGCCACCCGTAA	cce7424 m C	TCTTT A CTCTT A A CT		//cmaccome	CCCACTCCC	CACCATCCCC							Camme
cceOffB m AATTCTAACATGAAGCACTTAAATAAA/ <u>TTAGGGTGGGCAATGCCCACCTTA</u> GCAAGATAA AG <u>TAAGGTGGGCAATGCCCCACCTAC//AAACTATTAACTACTGCTACAAATAAATGC</u> alr4734 m GAAGCCGTAACGGTTAAATATAGTT/ <u>GTAGGTGGGCGGGGGGGGGGGGGGGGGGGGGGGGGGGGG</u>		TOTTINCIGITANC:	L'AAAAAGTAAT,	//GTAGGGTG	AGGCWCI GCC	CACCALCECE	CACAAA I	AAAAGGTGGG	CAGTGCCCA	CCTAC//GG	TTGTAATTAA	AAAAAAATA	GALLI
air4/34 m GAAGGCGTAACGGTTTAAATATAGTT//GTAGGTGGGGGGGGGG	cceOrfA rn A	TCGTAGTTAAAGACO	FAAAAAGTAAT, GAATGAGGATT,	// <mark>GTAGGGT</mark> G	GGCACTGCC	CACCTTAACC	ATTGCT 2	TAATGGTGGG	CAGTGCCCA CAATGCCCA	CCCTAC//GG	TTGTAATTAA AACTTAAAAT	GTAAAAAAAAA GTAATCCAAC	TGGTG
CCEUILD III GGTTTTUATUAATAUAATAUAATAUATTAUAAAAAU//GTAGGTGGGCAAATACTGTTTTATATT CTAAACTUTGGUTTTGGUUTACC//TTTTATTTAACTTCTATAACAATCTCA	cceOrfA rn A cceOrfB rn A	TCGTAGTTAAAGACO ATTCTAACATGAAGO	TAAAAAGTAAT, GAATGAGGATT, CACTTAAATAA,	//GTAGGGTG //GTAGGGTG //GTAGGGTG	GGCATTGCC GGCATTGCC GGCAATGCC	CACCTTAACC CACCTTAGCA	ATTGCT A AGATAA A	AAAAGGTGGG TAATGGTGGG GTAAGGTGGG	CAGTGCCCA CAATGCCCA CATTGCCCA	CCCTAC//GG CCCTAC//TI CCCTAC//AA	TTGTAATTAA AACTTAAAAT ACTATTAACT	GTAAAAAAAAAA GTAATCCAAC ATCTTGAATA	TGGTG
	cceOrfA rn A cceOrfB rn A alr4734 rn G	TCGTAGTTAAAGACO ATTCTAACATGAAGO AAGGCGTAACGGTT	TAAAAAGTAAT, GAATGAGGATT, CACTTAAATAA, TAAATATAGTT,	//GTAGGGTG //GTAGGGTG //GTAGGGTG	GGCATTGCC GGCATTGCC GGGCAATGCC GGGTGGAGGC	CACCTTAACC CACCTTAGCA GCGTGAGAGAGA	ATTGCT A AGATAA A AATTAA	AAAAGGTGGG TAATGGTGGG GTAAGGTGGG CTACACTTCG	CAGTGCCCA CAATGCCCA CATTGCCCA TTCCACCCA	CCCTAC//GG CCCTAC//TT CCCTAC//AA ACCTAC//TA	TTGTAATTAA AACTTAAAAT ACTATTAACT ACTAACTGCG	GTAAAAAAAAA GTAATCCAAC ATCTTGAATA ACTGCCTCAC	TGGTG AATCC

FIG. 6. Anabaena sp. RAYTs. Amino acid sequence (A) and nucleotide flanking regions (B) of IS(alr1015) and related ISs. Amino acid residues marked with an asterisk (\*) are within the catalytic region of the protein, residues marked with a vertical line (|) below the sequence are conserved in tyrosine transposases called RAYTs, and residues marked with black triangles are conserved both in RAYTs and in IS200 transposases (22). In panel B, IRs are underlined. //, the end of commonality.

initiation of predicted proteins (Fig. 7C). Perfect or nearly perfect IRs (Table 1) form the outer limits of the ISs and, with one exception, are flanked by perfect, 8-bp DRs.

**IS982 family.** Seven IS982 family members found by ISfinder and denoted ISNsp1s (cluster 1 in Table 1; Fig. 1D) are identical to each other or nearly so. Alr0590, in cluster 2, more closely resembles two *Gloeobacter violaceus* ORFs, Glr0150 and Glr0155, but lacks over 90 of their N-terminal amino acids, and Asl0588 resembles an even smaller fragment. Cluster 3 comprises Alr4082 and All8559, which are unusual in lacking IRs where they diverge in sequence, and Asr7385, a short, C-terminal fragment (see Fig. S7, cluster 3, in the supplemental material).

**IS***Azo13* **family.** Except for details of their IRs, IS(*alr7562*) and IS(*all8069*) are identical and are flanked by 3-bp DRs (Table 1; Fig. 1E). IS(*alr7562*) transposed within the 3' end of *all7563* in IS(*all7564*, *all7562*) (see Fig. S5, cluster 5, and S8, cluster 1, in the supplemental material). The nucleotide sequences of IS(*alr8019*) and of homologous ISs from *N. punc-tiforme* are 94% identical, but whereas Alr8019 has a single, 410-amino-acid (aa) transposase ORF, the *N. punctiforme* ISs have two ORFs of 198 aa and 190 aa. The difference in ORF number is the result of the presence of a T at position 696 in *alr8019* (Fig. 7D) and its absence from the *N. punctiforme* sequences, leading soon thereafter to a nonsense codon at position 710. Between those positions is the sequence A AAA

AAA AAG that is known to elicit translational frameshifting (32). The predicted amino acid sequence of All2145 is identical to that of Alr8019 until aa 300, and then diverges greatly (see Fig. S8, cluster 2, in the supplemental material).

**ISL3 family.** Both clusters of *Anabaena* sp. ISs within the ISL3 family have homologs in *N. punctiforme* (Table 1 and Fig. 1F; see Fig. S9 in the supplemental material). The long, cluster 2 ORF in IS(*alr8016*) is closely related to Alr7386C, the protein predicted when IS(*all7002, all7001*) and IS(*all7005, all7004*) (Fig. 5) are computationally removed from IS(*alr7386'*, etc.). Hypothetical proteins Alr7007, within IS(*alr7386'*, etc.) (Fig. 5), and Alr8017 in IS(*alr8016, alr8017*), respectively, are identical (data not shown). All *Anabaena* sp. sequences in the ISL3 family form extensive IRs flanked by perfect or imperfect DRs (Table 1).

Other transposases and possible transposases. Alr7163 and its nearly identical *A. variabilis* homolog, AvaB0061, are weakly similar to an ORF within, but not required for transposition of, IS493 (3), and evident IRs or DRs were not found (data not shown). All8065 and All8064 are found by BLASTp against the ISfinder database to be members of the ISAs1 family. Comparison with transposase genes of *Cyanothece* spp. and other cyanobacteria indicates that they belong to a single IS (data not shown), but their ends remain undetermined.

**ORFs interrupted, or otherwise affected, by ISs.** Once the presumptive ends and DRs of an IS were determined, the IS and one copy of directly repeated DNA were computationally

deleted. There were reciprocal reasons to identify proteins whose genes are mutated by ISs: first, to determine what functions may have been lost upon inactivation of genes by ISs and, second (see, e.g., Fig. 8D), to test our interpretations concerning the ends of ISs and the lengths of their flanking DRs. If those assessments were inaccurate, BLAST analysis (2) would have been expected to show gaps in the subject or the query when a sequence from which an IS was computationally removed was compared with its homologs. An error of 3n + 1 or 3n + 2 bp (*n*, an integer) could also have destroyed the reading frame of the predicted protein.

Of ca. 150 transposase ORFs in *Anabaena* sp., 120 were associated with 77 ISs that could be precisely removed computationally, plus 12 ISs whose DRs were imperfect. Four MITEs could also be removed precisely. The pretransposition form of each of these regions of the genome was then examined for an ORF that might have been interrupted by the IS. Of these ISs and MITEs, 17 are between convergent ORFs (data not shown) and so may have no significant physiological effect, whereas 8 are between divergent ORFs and 25 are between parallel ORFs (data not shown) and so have the possibility of affecting promotion of downstream ORFs. IS-(*all7002, all7001*), IS(*all7005, all7004*), IS(*alr7386'*, etc.) (Fig. 5), and IS(*all7562*) are within other ISs (see above), and IS(*alr1332*), IS(*all1608*), and IS(*alr7349*) were truncated by MITEa, IS(*alr1609*), and IS(*alr7350*), respectively.

One IS is present within one of four copies of 23S rRNA, and at least 29 other ISs appear to be inserted within presumptive, protein-encoding genes (Table 2). Presumptive proteins whose genes are interrupted (Table 2; see Table S1 in the supplemental material) include the following: a serine-threonine kinase, encoded by a fusion of parts of plasmid- $\alpha$  ORFs asr7230 and alr7232, that has nearly full-length orthologs in A. variabilis, Nodularia sp., and N. punctiforme; two-component histidine kinases mutated by IS(all3986) and IS(alr4104); an acetyltransferase and a peptidase interrupted by IS(all7115) and IS(all7302), respectively; an 88-aa DNA-binding protein with a PIN (PilT N terminus) domain, COG55673 (18), interrupted by IS(alr1858, alr1859); a type I restriction modification system DNA specificity subunit-of which Anabaena variabilis has a full-length homolog, Ava3267-mutated by IS(all3624); and a member of the Bpu10I (CCTNAGC) restriction enzyme superfamily (19), interrupted by IS(all4817, all4816).

In some instances in which a gene annotated as a "transposase" was near a short (50 to 99 aa) ORF annotated "asl..." or "asr..." (15), the short ORF comprised a genomic sequence, as it existed prior to insertion of the IS, and a sequence from within an end of the IS that lacked an intervening, in-frame stop codon. IS1594 thus contributed, in addition to its transposase gene, large parts of ORFs asl0305 and asl1098. More generally, IS1594 can provide a 66-aa termination for an ORF (if in the correct reading frame) initiated outside it and 49-aa initiation, starting with a GTG (preceded by what may function as a ribosome binding site), for an ORF extending outwards from it. Similarly, MITEb and MITEe provided more than half of the ORFs asl2519 and asl7509 and, when removed computationally, showed new, 66- and 67-aa ORFs, respectively. These, however, did not appear to be genes; i.e., homologs were sought but not found. IS891 and IS(alr1609) can provide

51- and 54-aa extensions, respectively, for ORFs initiated from outside.

#### DISCUSSION

Numerous Anabaena sp. ISs differ extensively from others in the same family. Those that are conspicuously truncated or whose transposase gene is interrupted are evidently simply inactivated. However, we believe that many others are not necessarily inactive but, rather, have evolved exogenously and entered separately. Specifically, Fig. 1 shows that for ISs in six families, Anabaena sp. ISs in clusters other than cluster 1 are much less similar to ISs in cluster 1 than to ISs found in other cyanobacteria. The comparisons depended on the availability, through genomic sequencing, of data on the ISs in numerous other cyanobacterial strains. Noncyanobacterial strains were not excluded. However, because transposases need to recognize the ends of their ISs, usually only very closely related transposase sequences were helpful in identifying the ends of their ISs. In practice, such protein sequences were nearly always found in cyanobacteria. When there were very few Anabaena sp. ISs in a cluster, only distantly related to any other IS in Anabaena sp., the ability to recognize the ends of ISs, or to increase the relative certainty of having identified such ends, often depended on knowing the sequences of related ISs in other organisms. Table 1 (footnote a) lists ISs at least one of whose ends could not be determined with assuredness. The absence of IRs and DRs enhances the difficulty of identifying the ends of IS891-related ISs, helping to explain why those ISs are prominent in this list. It is likely that in some instances, an end has vanished. However, it also appears likely that as more ISs are identified, the ends and transposases of some will match those in footnote a of Table 1, permitting analysis of those ISs and their effect on the genome. That is why we expect that further sequencing will help to extend the findings of this study to some of the unresolved ISs of Anabaena sp., as well as to other organisms.

There is plentiful precedent for translational frameshifting within portions of an IS (10), often in poly(A) sequences such as in A AAA AAG (32). It could be considered curious and remarkable that many of the differences found in this study that lead to there being different numbers of ORFs within IS895 (Fig. 7A), IS891-related cluster 2 (Fig. 7B), and IS892 (Fig. 7C) are within such sequences. Figure 7D, in which the mutated position is adjacent to a poly(A) sequence, is a similar example in the ISAzo13 family. These indel mutations may have survived because translational frameshifting could allow those mutants to generate intact transposases, even if at different translational rates than if frameshifting were not required. In addition, frameshift mutations can revert (29). Therefore, we suggest that at least some of the variant forms observed are genetically interconvertible versions that may be differentially capable of transposition. Too high a transposition rate might destroy a host; too low a transposition rate might reduce the competitiveness of an IS. We therefore conjecture that the role of poly(A) series within ISs is to provide a spectrum of functional genes that differ in their rates of producing transposase so as to assist their adaptation to a particular host and colonization of new hosts.

Figure 2E presents evidence that MITEc has transposed.

## Α

alr4628 rn alr4628 alr0552 rn alr0552 alr1726 rn alr1726	AGCCGATG AGCCGATG AGCCGATG AGCCGATG AGCCGATG AGCCGATG	AGAATGCTC AGAATGCTC AGAATGCTC AGAATGCTC AGAATGCTC AGAATGCTC 580	STTGCAGA STTGCAGA STTGCAGA STTGCAGA STTGCAGA STTGCAGA 590	ААААААА ААААААА АААААААА ААААААААА ААААА	-GAGATTGAA -GAGATTGAA AGAGATTGAA AGAGATTGAA -GAGATTGAA -GAGATTGA * 610	AACATTGCTGG AACATTGCTGG AACATTGCTGG AACATTGCTGG AACATTGCTGG 	SCAAATAACO SCAAATAACO SCAAATAACO SCAAATAACO SCAAATAACO SCAAATAACO SCAAATAACO	CGAGAAGAAA CGAGAAGAAA CGAGAAGAAA CGAGAAGAAA CGAGAAGAAA CGAGAAGAAA A 64(	TCGAAACAG( TCGAAACAG( TCGAAACAG( TCGAAACAG( TCGAAACAG( )	GAAAGTTAA GAAAGTTAA GAAAGTTAA GAAAGTTAA GAAAGTTAA 50	GAGTAT GAGTAT GAGTAT GAGTAT GAGTAT	IGTTAATTG IGTTAATTG IGTTAATTG IGTTAATTG 670	ATGAGTGT ATGAGTGT ATGAGTGT ATGAGTGT ATGAGTGT 680
													.     .
	.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0		1.1	1.2 kb
alr0552 🗖 alr1726 🗖						_							alr0553 alr1727
В	-	0.1	0.2	 0.3	0.4	0.5 0	++ 0.6 0.7	7 0.8	0.9	1.0	1.1	 1.2	++++ 1.3kb
	all all	4465 0315							all03	14 . ,			
all0315/4 all0315 all4465 r all4465	4rn gttta gttta gttta rn gttta gttta	390 	400 	410      <b>TGAATATA</b>   <b>TGAATATA</b>   <b>TGAATATA</b>	420 AAACCTCTGGA AAACCTCTGGA AAACCTCTGGA	430     ATGGTCACTTT ATGGTCACTTT ATGGTCACTTT ATGGTCACTTT	440   TGAAACTAG <i>X</i> TGAAACTAG <i>X</i> TGAAACTAG <i>X</i>	450 AAAGCAAATCAC AAAGCAAATCAC AAAGCAAATCAC AAAGCAAATCAC	460 TTTTACTGAC TTTTACTGAC TTTTACTGAC	470    2aaaaaaggt 2aaaaaaggt 2aaaaaggt	480   ATTGGCA ATTGGCA ATTGGCA	490 	500 
Ali0315/ Ali0315/ Ali4465	 314 vpgkk 314 vpgkk vpgkk	100   GYPQFKK-TT GYPQFKK-NI GYPQFKKNNI	110    TVOLNIKPL csveyktsg RSVEYKTSG	120 .   DGHFLKLES wslsetrkq WSLSETRKQ	130    <u>KSLLLTKKVI</u> gitftdkkgig 21TFTDKKGIG	140 	150 	160    vrlvrradgy	170 	180 	1   zigldvg TIGLDVG	90    Lkefytdsng LKEFYTDSNG	 200 .   ghtepnpk SHTEPNPK
Alio315/3 Ali4465	21(   314 fyrtg FYRTG	) 22   ekrlkfrqrn EKRLKFRORF	20    rvsrhqkgs RVSRH <u>O</u> KGS	230 .   tnrkkalnk TNRKKA <u>L</u> NK	240 	250 	260 	270 	280 	290 	300 <sup>†</sup>   wleyfgw wLEYFGW	) 310    KFGRITVAV KFGRITVAV	320 .     NPAYTTS NPAYTTS
C all7268 all7268 all7004 all7004 alr8566 alr8566 asr8502 alr8502	820 II GAGGCG TM GAGGCG GAGGCG TM GAGGCG GAGGCG I	830 GATAGAAAT GATAGAAAT GATAGAAAT GATAGACAT GATAGACAT GATAGACAT AT	840	850 GAATTGAAA GAATTGAAA GAATTGAAA GGAATTAAA GGAATTAAA GGAATTAAA GGAATTAAA	860 	870 TAGAGGATTAC TAGAGGATTAC TAGAGGATTAC TAGAGGATTAC TAGAGGATTAC TAGAGGATTAC	880   . GAGAAATTG/ GAGAAATTG/ GAGACATTG/ GAGACATTG/	890    AACGTAGTGTT AACGTAGTGTT AACGTAGTGTT AACGTAGTGTT	900 	910 	TGGTGAC TGGTGAC - GTGAC TGGCAAC TGGCAAC	920 .II TATTATT TATTATT TATTATT TATTATT TATTATC TATTATC	
	820	830	840	.11 850	860	 870	880	890	·· ··· ·· 900	. 910		.     920	
0.1 0	- ).2 0.3	0.4	0.5	0.6	0.7 0.	.8 0.9	1.0	1.1 1.2	2 1.3	1.4	1.5	1.2 kb	
all7268 all7005 alr8566					asr85	501						e e	III7004 IIr8502
0.1 0	- ).2 0.3	0.4	0.5	0.6	0.7 0.	.8 0.9	1.0	1.1 1.2	2 1.3	1.4	1.5	1.6 kb	
		al al al al al al al al al al al al	17268 1 17268 rn 1 17005 - 17004 rn 1 17004 rn 17004 1 17004 1 17005 - 17005 -	1050  Iggaaaaaa Iggaaaaaa Iggaaaaaa Iggaaaaaa	1060  Gaagtgettt. Gaagtgettt. Gaagtgettt. Aagtgettt. Aagtgettt.	1070 ACCACCACCTT ACCACCACCTT ACCACCACCTT ACCACCACCTT ACCACCACCTT ACCACCACCTT	1080 in TAGTTAACC' TAGTTAACC' TAGTTAACC' TAGTTAACC' TAATCAACC'	1090 талаласасстт галаласасстт галаласасстт галаласасстт галаласасстт галаласастт галаласастт галаласастт	1100  CTAGCTAAGG CTAGCTAAGG CTAGCTAAGG CTAGCTAAGG ATAGCTAAAG	.   . GGAT GGAT GGAT GGAT GGAT			
D all2145 rn all2145 alr8019 rn alr8019 np rn np	6   ACGAACCC ACGAACCC ACGAACCC ACGAACCC ACGAACCC	80 GTAATTTCA GTAATTTCA GTAATTTCA GTAATTTCA GTAATTTCA GTAATTTCA	690 	700    адаладада адалададада адалададада адалададада адалададада адалададада	710    Gaattaattg Gaattaattg Gaattaattg Gaattaattg Gaattaattg Gaattaattg	720    Gagattttaaj Gagattttaaj Gagattttaaj Gagattttaaj	730 AAATTCCGGA AAATTCCGGA AAATTCCGGA AAATTCCGGA	740   . ACCGAGTGGTG ACCGAGTGGTG ACCGAGTGGTG ACCGAGTGGTG ACCGAGTGGTG	750 	760    CGGCCAATT CGGCCAATT CGGCCAATT CGGCCAATT CAGCCAATT	770   GAGGTGAJ GAGGTGAJ GAGGTGAJ GAGGTGAJ	780 	790 ATTTTGTTG ATTTTGTTG ATTTTGTTG ATTTTGTTG ATTTTGTTG ATTTTGTTG

Δ	Left fla	nking region//Left	IR		Ri	<i>ght IR</i> //Right fl	anking region
<i>/ \</i>	50 60	70 80	90	1250 1260	1270	1280 1290	
1.alr0552 r 2.alr0552 r 3.alr0552 r 4.alr0552 r 5.alr0552 r	I TGATAGTTGTATAT n TGATAGTTGTATAT n TGATAGTTGTATAT n TGATAGTTGTATAT n TGATAGTTGTATAT	AA// <u>TATAGGAATCCTAT</u> AA <u>T</u> // <u>TAGGAATCCTAT</u> AATA// <u>TAGGAATCCTAT</u> AATA// <u>AGGAATCCTAT</u> AA <u>TAT</u> // <u>AGGAATCCTAT</u>	TGATTTGTGAACA CT TGATTTGTGAACA CT TGATTTGTGAACA CT TGATTTGTGAACA CT TGATTTGTGAACA CT TGATTTGTGAACA CT	TATTCAAAAATCAAATAA TATTCAAAAATCAAATAA TATTCAAAAAATCAAATAA TATTCAAAAAATCAAATAA TATTCAAAAATCAAATAA	GGATTCCTATA//O GGATTCCTAT//A GGATTCCTA//TA GGATTCCT//ATA GGATTCC//TATA	 CTAAATATTCAGTATT CTAAATATTCAGTATT CTAAATATTCAGTATT CTAAATATTCAGTATT CTAAATATTCAGTATT	TTA IR, no DR TTA IR, no DR TTA IR, 2-bp DRs TTA IR, 0 DR TTA IR, 4-bp DRs
<b>B</b> presump	tive //IS// and r	ight hand copy of I	DR removed at /				
1. minus IS 2. minus IS 3. minus IS 4. minus IS 5. minus IS	; there is no DR ; there is no DR minus 1 copy of ; there is no DR minus 1 copy of	ACGTAAAATGATAG ACGTAAAATGATAG DR ACGTAAAATGATAG ACGTAAAATGATAG DR ACGTAAAATGATAG	TGTATATAA/CTAAAT TGTATATAA <u>T</u> /ACTAA TGTATATAA <u>TA</u> /CTAA TGTATATAA <u>TA</u> /ATA TGTATATAA <u>TAT</u> /ATA	ATTCAGTATTTTACTGA ATATTCAGTATTTTACT( ATATTCAGTATTTTACT CTAAATATTCAGTATTTT AAATATTCAGTATTTTA	G 0 bp GAG 2 bp GAG 2 bp TACTGAG 6 bp CTGAG 4 bp	introduced introduced introduced introduced introduced	
C Anab: Avar:98478	cagcccagtatctacta             cagcccagtatttacta	gaattaaccctcactggg                  gaattaaccctcactggg	gggttgaaaaagttcgc                 ggttgaaaagtttgc	<b>TA</b> gttattttgacctca                  cagttattttgacctca	cagcaatgtcaaco               aagcaatgtcaac	ccgaagtttatataca               tagaagtttatataca	agcattgacca           agcattgacca 98586
<b>D</b> Query 227 Sbjct 227	FVNKMTSAITAQYLLE FVNKMTSAI AQYLLE FVNKMTSAIAAQYLLE	LTLTGGLKKF/ASYFDLT/ LTLTGGLKKF-ASYFDL / LTLTGGLKKF-ASYFDLK/	MSTRSLYTSIDQLKKY MSTRSLYTSIDQLKKY MSTRSLYTSIDQLKKY	YILQTNYHK 285 YILQTNYHK YILQTNYHK 285			
Query: ag JI Sbjct ag A.	ctatttaacagataaaa              ctatttaacaggtgaac var.396110	attttctcttgttttctgf                  gttttccctgattgtctgf	aaattotgtgtattta                   aatttotgtatattta	ttaggttta <b>TA</b> gaatata                       tcaaatttt <b>TA</b> ggacata	aagactttgattt              aaaatcttgattt	ttacagtatttttaca                 ttacggtgtttttaca	gaatttattcgatct               gaatttattcgatct 395996
F	50	60 70	80 90	1250	1260 1:	270 1280	1290
alr0552 rn alr1726 rn alr1853 rn al11972 rn al12066 rn alr2773 rn alr4628 rn al14867 rn	АСБТААААТGАТАСТТ ААТТСТСТСТАТАТАТТ ТТТАССААСААААСАТ ТТСТАСАТТТААСТТ ТТААААААСАТАААТТ АТТТААТGТАТААGT СТАGTСТАGGAGAGCC	GTATATAATA//TAGGAAT TAGGTTATA//TAGGAAT ATTCCTACTA//TAGGAAT GCCAGAGATA//TAGGAAT ATTCCAACTA//TAGGAAT ATTACAATCTA//TAGGAAT TATGAATCTA//TAGGAAT	CCTATTTGATTTGTGA CCTATTTGATTTGTGA CCTATTTGATTTG	I   ACA CTTATTCAAAAI   ACA CTTATTCAAAAI	ATCAAATAGGATT ATCAAATAGGATT ATCAAATAGGATT ATCAAATAGGATT ATCAAATAGGATT ATCAAATAGGATT ATCAAATAGGATT ATCAAATAGGATT	CCTA//TACTAAATAT CCTA//TAGAATATAA CCTA//TAGAATATAA CCTA//TAGTACTCATG CCTA//TAGTCCTATG CCTA//TAGACGGTA CCTA//TATTATATA CCTA//TATTATCTCGA CCTA//TAGCCCTATA	CACTATTTACTAC GACTTTGATTTTAC TTGTTAGGAGTAGGT AACCATTAATAGAAA GTAGTTAGTAGGGTG AATACTAGAAAATAC TATTTTATTT

FIG. 8. Palindromes at the ends of IRs of copies of IS895-related ISs. (A) If an IS ends in, or is flanked by, a palindrome (here, TA or TATA, and underlined), that palindrome can form part of the IRs (interpretations 1 to 4), DRs (3 or 5), or both (3), resulting, as shown in panel B, in different frameshifts when the IS and one copy of a DR are computationally removed. Further information is required to determine which is correct in a particular case. (C to E) Evidence that IS(*all7564*, *all7563*) in cluster 5 (C and D) and IS(*alr1726*, *alr1727*) in cluster 1 (E) generate a TA direct repeat (capitalized and underlined) upon insertion, by comparison with homologous nucleotide (C and E) and protein-encoding (D) regions in *A. variabilis* (Subject). (F) That interpretation (A3, B3) is here shown applied to all members of cluster 1.

The ends of its IRs are closely related to those of IS(*alr1332*) and its *Synechococcus* sp. relatives. Therefore, the transposase of IS(*alr1332*)—before it lost its R end—might have catalyzed transposition of those MITEs in *Anabaena* sp. If MITEa trans-

posed into its now fragmented "parent," the MITE may have transposed before the R end of IS(*alr1332*) was repositioned or lost, in which case the small ORF that is intact in three of the MITEs (Fig. 2C) may have no catalytic significance. If so, and

FIG. 7. Frameshifting resulting from indel mutations. (A) IS895s. The partial sequence and the bars that represent IS(*alr0552*, *alr0553*) also represent IS(*alr1853*, *alr1854*), IS(*all1972*, *all1971*), IS(*all2067*, *all2066*), IS(*alr2773*, *alr2774*), and IS(*all4868*, *all4867*). The poly(A) region, positions 594 through 603, determines how long an ORF continues. The ORFs end (\*) at position 658 if all 10 As remain, at position "611" if 1 A is removed, and (not shown) 9 bp from the R end of the IS if 2 As are removed. (B) IS891-related cluster 2. The partial sequence and the bar that represent IS(*all4465*) also represent IS(*all7325*). Due to a deleted nucleotide at position ca. 392, All0315 terminates at amino acid position 136 (\*), whereas All4465 continues further. Lowercase black letters in the bottom parts of panel B represent the hypothetical amino acid sequence in the absence of the single-nucleotide deletion. (C) IS892, aka IS(*all7268*), and related ISs. The partial sequence and the bar that represent IS(*all7323*) and IS(*alr8510*). With a single base pair absent from the poly(A) region of positions 850 through 859, the ORF continues to 18 bp before the R end of the IS (not shown). The partial sequence and the bars that represent IS(*all7005*, *all7004*) also represent IS(*all7106*, *all7112*, *all7111*), IS(*all7178*, *all7177*), IS(*all7303*, *all7302*), and IS(*all7376*, *all7375*). With no deletions in the aforementioned poly(A) region, the first ORF ends at position 1064. With the G lost at position 1053 within the A-rich region of positions 1047 to 1055 (bottom part of panel C), *asr8501* ends at position 1076. (D) IS*Azo13*-family, cluster 2. Deletion of a T at position 696 of *N. punctiforme* ISs (represented by np) adjacent to a poly(A) region causes truncation of the np ORFs at position 712. However, as in panels A to C, ribosomal frameshifting in the immediately succeeding poly(A) sequence may allow continued translation of the sequence.

TABLE 2. ORFs and other loci affected by inse	rtion of I	Ss
---	------------	----

Family/group/CL <sup>a</sup>	Comment
IS4/ISPepr1/CL ISAcma11	IS( <i>all7115</i> ) interrupts a GNAT superfamily protein gene, with alignment of homologs Npun_R0240 and PCC7474 2037 across the mutation site.
IS5/IS1031/CL 1	IS( <i>alr4438</i> , <i>alr439</i> ) is within a questionably significant, 55-aa ORF. IS( <i>all4817</i> , <i>all4816</i> ) is within a Bpu10I RE Superfamily protein gene (19). IS( <i>alr5157</i> , <i>alr5158</i> ) is within a gene that fuses the start of <i>alr5156</i> and the end of <i>alr5159</i> and that has homologs in diverse cyanobacteria and other bacteria. BLASTn shows that IS( <i>all7002</i> , <i>all7001</i> ) can be computationally excised from a sequence closely matching that of <i>alr8016</i> . Computational removal of all of these ISs leads to no shift in the reading frame.
IS5/IS1031/CL 2	IS( <i>alr7025</i> ) revises the N terminus of a membrane protein presumptively encoded by <i>all7024</i> .
IS110/(-)/IS1594	Removing IS( <i>all0306</i> ) destroys <i>asl0305</i> , and removing IS( <i>all1099</i> ) destroys <i>asl1098</i> . IS( <i>alr3636</i> ) is inserted within an ORF that may encode a DNA mismatch repair protein.
IS200-IS605/IS1341/IS891	Each of IS( <i>all3986</i> ) and IS( <i>alr4104</i> ) splits a two-component His kinase family gene, and IS( <i>alr7231</i> ) splits a Ser/Thr kinase family gene.
IS200-IS605/IS1341/IS891-related CL 2	BLASTn localizes IS(all0315, all0314)-, IS(alr1157)-, and IS(all4465)-flanking regions between ava4874 and ava4875, ava4411 and ava4412, and ava3329 and ava3330, respectively.
IS200-IS605/IS1341/IS891-related CL 6	IS(all7085, alr7086) revises the 5' terminus of all7084.
IS630/(-)/IS895-related CL 1	IS( <i>alr1726</i> , <i>alr1727</i> ) is within a 102-aa ORF that may not be a gene. IS( <i>alr1853</i> , <i>alr1854</i> ) is within a 68-aa ORF that may encode a protein. Removal of IS( <i>all1972</i> , <i>all1971</i> ) extends the 5' end of <i>all1973</i> to 37 bp from <i>all1970</i> .
IS630/(-)/IS895-related CL 3	IS( <i>alr1858</i> , <i>alr1859</i> ) splits an 88-aa ORF that putatively encodes a protein with a PIN (PiIT N terminus) domain in diverse cyanobacteria and other bacteria.
IS630/(-)/IS895-related CL 5	BLASTn and BLASTp localize IS( <i>all7564</i> , <i>all7563</i> ) in an ORF that encodes an E1 enzyme (super)family protein.
IS892	IS( <i>all7178</i> , <i>all7177</i> ) is inserted within an unannotated, 417-bp ORF that overlaps <i>all7178</i> and has strong homologs, including a close match within <i>N. punctiforme</i> ORF <i>npR1908</i> . IS( <i>all7005</i> , <i>all7004</i> ), IS( <i>all7106</i> , <i>all7105</i> ), IS( <i>all7268</i> ), IS( <i>all7303</i> , <i>all7302</i> ), IS( <i>alr7323</i> ), and IS( <i>alr8510</i> ) are inserted in genes that, respectively, encode a transposase, and—presumptively—a phage resistance protein, a 315-aa ORF with a HAD domain, a member of the DUF 1392 superfamily, a 444-aa ORF with a peptidase domain, and a Mob-Pre plasmid recombination enzyme. Ends of all of these ISs were confirmed by alignment with homologs of proteins whose genes the ISs mutated. IS( <i>all7112</i> , <i>all7111</i> ) is inserted 19 bp 5' from <i>all7110</i> , presumably affecting its transcription.
IS982/(-)/CL 1	Removal of IS( <i>all2664</i> ) led to the finding of a 266-aa ORF (Table S1 in the supplemental material) that fuses the 3' end of <i>asr2665</i> to and through <i>asr2666</i> . IS( <i>all3624</i> ) is positioned within a fusion of <i>alr3623</i> and <i>alr3625</i> that encodes a type I restriction modification system DNA specificity subunit. Removing ISNsp1 (aka IS[ <i>alr1569</i> ]) extends <i>asr1570 5'</i> , predicting a DUF 196 superfamily protein. BLASTn and BLASTp show that ISNSp1 is inserted in a sequence that closely matches, across the mutation site, that of the start of <i>ava4176</i> .
IS982/(-)/CL 3	IS( <i>alr4082</i> ) sits in an ORF whose predicted product has tBLASTn scores of >6e-25 in other strains.
ISAzo13/(-)/CL ISNsp4	The sequence intercepted by IS( <i>alr8019</i> ) is highly similar to (89% amino acid identity predicted)
	but only 63% as long as <i>all</i> 7355, with alignment across the mutation site. A 426-bp, unannotated ORF that fuses the ends of <i>alr</i> 2144 and <i>all</i> 2145 is highly homologous to <i>A. variabilis</i> predicted protein AvaC0004. IS( <i>all</i> 2145) ends truncated 40 bp 3' from that 426-bp ORF.
ISL3/(-)/CL 1	IS( <i>alr2698</i> ) is inserted five codons before the end of <i>alr2697</i> .

<sup>a</sup> CL, cluster; -, not assigned by ISfinder.

because Alr5204 and All7115 appear to be structurally defective, no IS4-family IS of *Anabaena* sp. may remain active.

ISs, a major constituent in the genomes of many prokaryotes, account for ca. 2.4% of the protein-encoding genes in *Anabaena* sp. About one-third of the ISs herein elucidated interrupt protein-encoding genes; others, inserted between parallel or divergent ORFs, may also have significantly affected cellular metabolism (see, e.g., reference 16). Because mutations in essential genes or operons would be expected to be highly disadvantageous, if not lethal, one would not expect to find ISs within such genes or operons. Rather, they may be present in genes that are normally inessential but might be helpful, or even essential, under specific conditions. Such conditions might include a need for flotation by means of gas vacuoles or a need for the differentiation of akinetes, a form of sporulation. *Anabaena* sp. is not known to form gas vacuoles or akinetes, but genes required for gas vacuole formation or selectively expressed in akinetes are present (20, 34). Although no ISs or known MITEs (33; also this paper) are found near those genes, it remains possible that a regulatory gene mutated by an IS may have been required for activation of those genes. The fact that a copy of IS1594 was found within a 23S rRNA may be understood by the fact that it is only one of four copies of such a gene. We suggest that removal of that copy of IS1594 by recombinant DNA genetic manipulation (6) might enable Anabaena sp. to grow more rapidly. Had a type I restriction modification system gene similar to A. variabilis gene ava3267 and a Bpu10I restriction enzyme superfamily gene not been mutated, as they were by IS(all3624) and IS(all4817, all4816), respectively, gene transfer to Anabaena sp. might not have been achieved or might have been achieved only with greater difficulty or lower frequency (13). Conversely, inactivation of ava3267 might enhance gene transfer to A. variabilis. We also found, repeatedly, that certain "hypothetical" or "unknown"

ORFs resulted from ORFs at the ends of ISs or of MITEs fusing with short, intergenic ORFs from *Anabaena* sp.

It is unclear why IRs are sometimes flanked by imperfect DRs, especially when other ISs in a cluster show perfect DRs. Perhaps two copies of the IS inserted near each other, in the same orientation, and then underwent homologous recombination, deleting the intervening sequence. Another possibility is that second-strand synthesis of staggered repeats may have been inaccurate.

Four instances were noted, above (see, e.g., Fig. 5) in which ISs are present within ISs, plus instances in which ISs truncated additional ISs. If, as seems highly likely, ISs have evolved as prospective hosts evolved to protect against them, an IS that is found within, or has truncated, the transposase of another IS either arose later than that which it intercepted or can be defined as contemporaneous with the latter. By examining a large collection of nested ISs, one might be able to determine a relative temporal order in which ISs evolved and could compare that order with the changes in sequence taking place in the ISs involved. Alternatively, such a collection might show that IS evolution cannot be ordered in that way.

#### ACKNOWLEDGMENTS

We are grateful to Jeff Elhai (Virginia Commonwealth University) for an extensive, helpful critique.

This work was supported by the Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy grant DOE FG02-91ER20021.

#### REFERENCES

- Alam, J., J. M. Vrba, Y. Cai, J. A. Martin, L. J. Weislo, and S. E. Curtis. 1991. Characterization of the IS895 family of insertion sequences from the cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. **173**:5778–5783.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.
- Baltz, R. H., D. R. Hahn, M. A. McHenney, and P. J. Solenberg. 1992. Transposition of Tn5096 and related transposons in *Streptomyces* species. Gene 115:61–65.
- Bancroft, I., and C. P. Wolk. 1989. Characterization of an insertion sequence (IS891) of novel structure from the cyanobacterium *Anabaena* sp. strain M-131. J. Bacteriol. 171:5949–5954.
- Bancroft, I., C. P. Wolk, and E. V. Oren. 1989. Physical and genetic maps of the genome of the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. 171:5940–5948.
- Borthakur, P. B., C. C. Orozco, S. S. Young-Robbins, R. Haselkorn, and S. M. Callahan. 2005. Inactivation of *patS* and *hetN* causes lethal levels of heterocyst differentiation in the filamentous cyanobacterium *Anabaena* sp. PCC 7120. Mol. Microbiol. 57:111–123.
- Cai, Y. 1991. Characterization of insertion sequence IS892 and related elements from the cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. 173:5771–5777.
- Cai, Y. 1991. Molecular genetic approaches towards the understanding of heterocyst differentiation and pattern formation in the cyanobacterium *Anabaena* sp. Ph.D. thesis. Michigan State University, East Lansing, MI.
- Cai, Y., and C. P. Wolk. 1990. Use of a conditionally lethal gene in *Anabaena* sp. strain PCC 7120 to select for double recombinants and to entrap insertion sequences. J. Bacteriol. 172:3138–3145.
- Chandler, M., and J. Mahillon. 2002. Insertion sequences revisited, p. 305– 366. In N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz (ed.), Mobile DNA II. ASM Press, Washington, DC.
- Chen, Y., F. Zhou, G. Li, and Y. Xu. 2009. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. Gene 436:1–7.
- Elhai, J., M. Kato, S. Cousins, P. Lindblad, and J. L. Costa. 2008. Very small mobile repeated elements in cyanobacterial genomes. Genome Res. 18: 1484–1499.

- Elhai, J., A. Vepritskiy, A. M. Muro-Pastor, E. Flores, and C. P. Wolk. 1997. Reduction of conjugal transfer efficiency by three restriction activities of *Anabaena* sp. strain PCC 7120. J. Bacteriol. 179:1998–2005.
- Galas, D. J., and M. Chandler. 1989. Bacterial insertion sequences, p. 109– 162. In D. E. Berg and M. M. Howe (ed.), Mobile DNA. American Society for Microbiology, Washington, DC.
- Kaneko, T., Y. Nakamura, C. P. Wolk, T. Kuritz, S. Sasamoto, A. Watanabe, M. Iriguchi, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, N. Nakazaki, S. Shimpo, M. Sugimoto, M. Takazawa, M. Yamada, M. Yasuda, and S. Tabata. 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8:205–213.
- Luque, I., A. Andújar, L. Jia, G. Zabulon, N. Tandeau de Marsac, E. Flores, and J. Houmard. 2006. Regulated expression of glutamyl-tRNA synthetase is directed by a mobile genetic element in the cyanobacterium *Tolypothrix* sp. PCC 7601. Mol. Microbiol. 60:1276–1288.
- Mahillon, J., and M. Chandler. 1998. Insertion sequences. Microbiol. Mol. Biol. Rev. 62:725–774.
- 18. Marchler-Bauer, A., J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, and S. H. Bryant. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37:D205–D210.
- Matveyev, A. V., K. T. Young, A. Meng, and J. Elhai. 2001. DNA methyltransferases of the cyanobacterium *Anabaena* PCC 7120. Nucleic Acids Res. 29:1491–1506.
- Mlouka, A., K. Comte, A. M. Castets, C. Bouchier, and N. Tandeau de Marsac. 2004. The gas vesicle gene cluster from *Microcystis aeruginosa* and DNA rearrangements that lead to loss of cell buoyancy. J. Bacteriol. 186: 2355–2365.
- Muro-Pastor, A. M., T. Kuritz, E. Flores, A. Herrero, and C. P. Wolk. 1994. Transfer of a genetic marker from a megaplasmid of *Anabaena* sp. strain PCC 7120 to a megaplasmid of a different *Anabaena* strain. J. Bacteriol. 176:1093–1098.
- Nunvar, J., T. Huckova, and I. Licha. 2010. Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. BMC Genomics 11:44.
- 23. Ohmori, M., M. Ikeuchi, N. Sato, P. Wolk, T. Kaneko, T. Ogawa, M. Kanehisa, S. Goto, S. Kawashima, S. Okamoto, H. Yoshimura, H. Katoh, T. Fujisawa, S. Ehira, A. Kamei, S. Yoshihara, R. Narikawa, and S. Tabata. 2001. Characterization of genes encoding multi-domain proteins in the genome of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8:271–284.
- 24. Pei, A., C. W. Nossa, P. Chokshi, M. J. Blaser, L. Yang, D. M. Rosmarin, and Z. Pei. 2009. Diversity of 23S rRNA genes within individual prokaryotic genomes. PLoS One 4:e5437.
- 25. Rippka, R., R. W. Castenholz, and M. Herdman. 2001. Subsection IV, p. 562–580. In D. R. Boone, R. W. Castenholz, and G. M. Garrity (ed.), Bergey's manual of systematic bacteriology, 2nd ed., vol. 1. The *Archaea* and the deeply branching and phototrophic *Bacteria*. Springer-Verlag, New York, NY.
- Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.
- Schwarz, R., and M. Dayhoff. 1979. Matrices for detecting distant relationships, p. 353–358. *In* M. Dayhoff (ed.), Atlas of protein sequences. National Biomedical Research Foundation, New York, NY.
- Siguier, P., J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34:D32–D36.
- Snyder, L., and W. Champness. 1997. Molecular genetics of bacteria. American Society for Microbiology, Washington, DC.
- Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24:1596–1599.
- Wolk, C. P., J. Elhai, T. Kuritz, and D. Holland. 1993. Amplified expression of a transcriptional pattern formed during development of *Anabaena*. Mol. Microbiol. 7:441–445.
- Zhou, F., V. Olman, and Y. Xu. 2008. Insertion sequences show diverse recent activities in cyanobacteria and archaea. BMC Genomics 9:36.
- Zhou, F., T. Tran, and Y. Xu. 2008. Nezha, a novel active miniature invertedrepeat transposable element in cyanobacteria. Biochem. Biophys. Res. Commun. 365:790–794.
- Zhou, R., and C. P. Wolk. 2002. Identification of an akinete marker gene in Anabaena variabilis. J. Bacteriol. 184:2529–2532.