JOURNAL OF BACTERIOLOGY, Oct. 2010, p. 5289–5303 Vol. 192, No. 20
0021-9193/10/$12.00 doi:10.1128/JB.00460-10

The Insertion Sequences of *Anabaena* sp. Strain PCC 7120 and Their Effects on Its Open Reading Frames▽†

C. Peter Wolk,[1,2]* Sigal Lechno-Yossef,[1]‡ and Karin M. Jäger[1]§

MSU-DOE Plant Research Laboratory[1] and Department of Plant Biology,[2] 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 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., IS*1594* or IS*5*, 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(*alr4628*) or IS(*all7002*, *all7001*). 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

* 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.

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

§ Present address: Radolfzeller Str. 78, D-78467, Konstanz, Germany.

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

▽ Published ahead of print on 23 July 2010.

TABLE 1. Clustered ISs of *Anabaena* sp.: how their ends were identified, selected characteristics, and related supplemental figures

Family/group[a]	CL and/or *Anabaena* sp. PCC 7120 CL members[c]	Other members of the CL[d]	L end (excluding variants)[e]	Inverted R end (excluding variants)[f]	E value of IRs[g]	Typical DR (no. of bp)	Tentative target and IS insertion point (/)[i]	Basis for identification of ends[j]	Supplemental figure reference
IS*4*/IS*50*	IS(*alr1332*), 5 MITEs	4 *Sy*7002	CTACG GTGTA CACAC AAGTC CAAGT AAAGT	CTACC GTGTA CACAC AAGTA CTATC TGCAA	5E−9	10 or 11	GC-rich	A, IRs, DRs	S1 CL 1
IS*4*/IS*4Sa*	IS(*alr5204*)	2 *Av*, 7 *Np*	CAGAA GTGTT GAATG TTAAG AAAAA GATGA GAAAA ATAGC	CAAAA ATGTT GAAAG CTGAT ACAAA ATTTT ACATA GTTAG	0.058	Imperfect 9	AT-rich	A, IRs	S1 CL 2
IS*4*/IS*Pepr1*	IS(*all7115*)	7 *Np*	CAATA CCTTA GCCAA AATAA GAGCA TAAAG AGGTA GGGCG	CAATA CCTCT GCCAA ATTAC GAGGG TTCAA CACCC GTAGA	0.015	6	Usually AT-rich	A, IRs, DRs, Table 2	S1 CL 3
IS*5*/IS*1031*	**CL 1**: IS(*all0016-15*), IS(*all2693-92*), IS(*alr3610-11*), IS(*all4400, all4399*), IS(*alr4438-39*), IS(*all4817-16*), IS(*alr5157-58*), IS(*all7002, all7001*)		GAGG CTATT TATAA AGTAA ATCTA AAGGA GAGCT ATCAG	GAGA CCATT TATAA AGTAA ATCTT TAGAC GACTA GACGA	6E−8	3	TWA (one TCA)	A, IRs, DRs, Table 2	S2 CL 1
IS*5*/IS*1031*	**CL 2**: IS(*alr7025*)	*Am*, 3 *Np*	GAGG GTGTT TGAAA AGTAG GGGAT GTTGT AAAAA AACTC CCTCG GTATA	GAGG ATGTT TGAAA AGTTA TAGGG AGTCA AAATT AAGCC AATCG CTTCA	6E−6	3	TWA	A, IRs, DRs	S2 CL 2
IS*110*/(−)	**IS*1594***: IS(*alr0249*), IS(*all0306*), IS(*all0732*), IS(*all1099*), IS(*alr1212*), IS(*all1986*), IS(*all2065*), IS(*alr3571*), IS(*alr3636*), IS(*all3682*), IS(*all3734*), IS(*all4756*)	*Np*	TGTAT ATTAA AAGAA GTGGT AGACC GTCGC	AGCGA CTGTC TTGAA AGTCA AGCGA TCGTT	NSS	0	CCT/AC, CC/TAC, or C/CTAC	A, K (rRNA), Fig. 1B	S3
IS*200*-IS*605*/IS*1341*	**IS*891***: IS(*all3986*), IS(*alr4104*), IS(*all5207*), IS(*alr7228*), IS(*alr7231*), IS(*all8010*)	*Ns*, 3 *Lyn*	GAGCC GTGAA GCGTA AAGCC CCCGT ATTTT	TTGAC ATCCT CCCCC GTTTA GAAAA CGGGG	0.032	0	TTAC/	K (see text), TGTCAA at R terminus, Table 2	S4 CL 1
IS*200*-IS*605*/IS*1341*	**IS*891*-related CL 2**: IS(*all0315-14*), IS(*alr1157*), IS(*all4465*), IS(*alr7325*)	*Av*	CAAGA AACTG GGTCT AAAGC CCCGT CCTTG	TTGAC ACTCT CCGCC CTATA AGTGC GGAGA	NSS	0	TTAC/	A, TGTCAA at R terminus, Table 2	S4 CL 2
IS*200*-IS*605*/IS*1341*	**IS*891*-related CL 3**: IS(*all2167, alr2168*), partial IS(*all1608*)	9 *Te*	CAAAA GAATG GGATA CAAGC CCCGT CGTTC TAGGA CGGCT	TTGAC ATACT CACCG ACCTA AAGGT GCGGT GATTC TTGAC	NSS	0	TGAC/	A	S4 CL 3
IS*200*-IS*605*/IS*1341*	**IS*891*-related CL 4**: IS(*alr1531*) (left end unclear)	*Av*, 7 *Te*	TGGTA AAATG TGAGG TATGG AAAAA GCCTA CCGCT ACCGA	TTGAC ATCCT CACCG CCCTG AAAGT GCGGT GATTC CTAAG	NSS	Unclear	Unclear	A, TGTCAA at R terminus, L terminus unclear	S4 CL 4
IS*200*-IS*605*/IS*1341*, (−)	**IS*891*-related CL 5**: IS(*all7148, alr7149*), IS(*all7008, alr7009*)	*Av*, *Cy*7424	AGTTT CTCAA AAATA TATTG ATGTT AGACG	TTGAC ACTCT CGCCG CTAAC CGCAA AGCAG	NSS	0	TTAC?	A, TGTCAA at R terminus	S4 CL 5
IS*200*-IS*605*/IS*608*	**IS*891*-related CL 6**: IS(*all3371, alr3372*), IS(*all7085, alr7086*) (approximately, respectively, IS*Nsp2* and IS*Nsp3* of ISfinder)	2 *Np*	GAGTC GTGAT GCGTA AAGCC CCCAA TTATG	TTGAG CCACT CCCCC GTTTT GAAAA CGGGG	NSS	0	TTAC	A, TCAA at R terminus	S4 CL 6

IS*200*-IS*605*[b]	IS(*alr1015*), IS(*alr4734*)	*Cy*8801, *Cy*0110, M*c;* IS*Ljo5*_ a1	GTAG**G GTGGG** **CAATG CCCAC** ***C***AAAA ATATT ATGTA AAAAT	GTAG**G GTGGG** **CATTG CCCAC** ***C***AATT ATCTC ATTAT GTAGT	4E−9	3		A, IRs	S4 CL 7
IS*630*/(−)	**IS*895***: IS(*alr0552-53*), IS(*alr1726-27*), IS(*alr1853-54*), IS(*all1972-71*), IS(*all2067-66*), IS(*alr2773-74*), IS(*alr4628*), IS(*all4868-67*)		TAGGA ATCCT ATTTG ATTTG TGAAC AAGAC CAAGA	TAGGA ATCCT ATTTG ATTTT TGAAT AAGTT CCGTA	8E−10	2	TA	A, IRs	S5 CL 1
IS*630*/(−)	**IS*895*-related CL 2**: IS(*asl1992*), IS(*asr3082*) (diverges from others close to R end)	*Av*, *Np*, 2 *Cw*, 4 *Am*	ACCAA TTTAA ATTAG AGACA GGGCA GATGA GGTAA	ACCAA ATTAA ATGGT GTTTA GGGCA GATAG GGCAT	0.003	2	ATAT	A, IRs	S5 CL 2
IS*630*/(−)	**IS*895*-related CL 3**: IS(*alr0018-19*), IS(*all0363-62*), IS(*alr1858-59*)	*Cw*, *Np*, *Cy*7425	TAGCG TTTAC CAGTA TAATG AAGTA CACTA ATTAA AATAA	TAGCG TTTCT CAGTC TGGTG AAGTA CAGTA ACAAG AATGG	0.004	2	TA	A, IRs, Table 2	S5 CL 3
IS*630*/(−)	**IS*895*-related CL 4**: IS(*alr5227-28*)	*Av*, 7 *Ns*	AGTAG GTAGG CACGA AAAAA CCAAA CTATG TGAAG ATAAG TAAAG ACGGA GAATA AATCT	AGTAG GTGGG TGGGA AAAGT CCCAA GTATG TAACG AAACA TTAAG TAATA GAATT AGAGT	0.011	2	TA	A, IRs	S5 CL 4
IS*630*/(−)	**IS*895*-related CL 5**: IS(*all7564-63*)	17 *Cw*	GTACA GGTCG GCGTA AATAA ACAGA CCATA	GTACA CCTCG GCGTA AATCA GCAGA CCATT	2E−6	2	CTAG	A, IRs, Table 2 (unchanged reading frame)	S5 CL 5
IS*892* (unclassified) [aka *IS*(*all7268*)][k]	**IS*892***: IS(*all7005-04*), IS(*all7106-05*), IS(*all7112-11*), IS(*all7178-77*), IS(*all7303-02*), IS(*alr7323*), IS(*all7376-75*), IS(*alr8510*), IS(*alr8566*, *asr8501*, *alr8502*)		CTAGC GTGGC AAAAC TTACT AGAGA GGAGC AGAGA TCCTG	CTAGC GTGGC AAAAC TTACT AGAGA AGACG ACTCT CTAGA	2E−12	8[h]	AT-rich	A, IRs, DRs, Table 2	S6 CL 1
IS*982*/(−)	**CL 1**: IS(*alr0999*), IS(*all2664*), IS(*alr2683*), IS(*alr2694*), IS(*alr3384*), IS(*all3624*) IS*Nsp1* [aka IS(*alr1569*)]		ACGTG ATGTG CGACT TATTG TTTCG TTACA CAATT GAGGT	ACGCC ATGTG CGACT TAATA TTCTG TAACA AGATC GTCGA	6E−5	mostly 6	AT-rich	A, IRs, DRs, Table 2	S7 CL 1
IS*982*/(−)	**CL 2**: IS(*asl0588*), IS(*alr0590*)	2 *Gvi*	ACGTG AGTTC GACGG GTTAA TTTAG GTGAA	ACGTG AGTTC GACGA ACTAA AAAAC AGCAG	2E−6	8 or less	AT-rich	A, IRs, DRs	S7 CL 2
IS*982*/(−)	**CL 3**: IS(*all8559*), IS(*alr4082*), IS(*asr7385*) (a fragment)	IS*Rmsp1*	ATTTA GGGTT TGTGC GAGCC AACTA TTTGA	TACGC CTTAT GTGAA TTAAG CCGGA GGATG	NSS	1?	AT-rich	A, Table 2	S7 CL 3
IS*Azo13*/(−)	**CL 1**: IS(*alr7562*), IS(*all8069*)	IS*Stau6*	GAGAA CTGCA CAGAA TGATT GATCC TATGA TCAGA GAAAG	GAGAA CTCCA CAAAA AAGAT GATCC AATAG CTATG CTGGT	0.015	3	AT-rich	A, IRs, DRs	S8 CL 1
IS*Azo13*/(−)	**CL 2**: IS*Nsp4* [aka IS(*alr8019*)], IS(*all2145*) (truncated at its R end); IS(*asr7385*) is an R-end fragment	8 *Np*	AGGCA TCATG TAAAA ATAAC TTGAA CGATT TACCG AATAG TTAGA	AGGAG TTATG TAAAA ATAAC CTGAA CAATT AAGTG CCTAC TTTGG	2E−6	3	TWA in AT-rich region	IRs, Table 2	S8 CL 2

Continued on following page

TABLE 1—*Continued*

Family/group[a]	CL and/or *Anabaena* sp. PCC 7120 CL members[c]	Other members of the CL[d]	L end (excluding variants)[e]	Inverted R end (excluding variants)[f]	E value of IRs[g]	Typical DR (no. of bp)	Tentative target and IS insertion point (/)[i]	Basis for identification of ends[j]	Supplemental figure reference
IS*L3*/(−)	**CL 1**: IS(*alr1609*), IS(*alr2698*), IS(*all7161*), IS(*alr7305*) (approximating IS*Asp1*), IS(*alr7349*) (truncated), IS(*alr7350*)	3 *Np*	GGTTC TTTCG GATAT TTTAT GGAGA AAGCA AAAAG TAATG AAAAT TAATG	GGTTC TTGCG CCTGT TTTAT GGAGA ATTAA TACTA AAGTG CCAGT TTAAT	E−4	Up to 8 perfect, often imperfect	AT-rich	A, IRs, DRs	S9 CL 1
IS*L3*/(−)	**CL 2**: IS(*alr7386'*, *alr7003'*, *asr7006; alr7007*), IS(*alr8016-17*)	2 *Np*	GGTTC TTGGC AACTT TTGGT GATCT TGGTT GGGGA AAGGC AGAAG GCAGG AGTCA GAAAG ATATA ATTGA	GGTTC TTGGC AACTT TTGGT GATCT TGGTT GGGGA AAGGC AGAGG GCAGA GGGCA GAGGG CAGAA GGCAG	5E−23	Up to 8 perfect, often imperfect	AT-rich	A, IRs, DRs	S9 CL 2

[a] ISs bearing the following transposase ORFs cannot (yet) be excised computationally: IS*5* family, *all2152*; IS*200*/IS*605* family, *alr1531*, *alr1685*, *alr2719*, *all4675*, *all5207* (see Fig. S4 in the supplemental material, cluster [CL] 1), *all7008* and *alr7009* (Fig. 4E and F), *alr7153* (perhaps in the IS*607* family), *all7158* (closely related to the IS*891* transposase gene), *all7245*, *asl7246*, *alr7329*, *all8070*, *alr8071*; and unclassified by ISfinder: *alr1015* and *alr4734* (see the text and Fig. S4, CL 7); IS*481* family, *all3630*; IS607 family, *asr7146*, *alr7147*, *asr7152*; IS630 family, *asl1657*, *alr1926*, *asr3082* (see Fig. S5, CL 3); IS*982* family, *asl0588* (very short, but retains IR and DR; see Fig. S7, CL 2); *asr7385* (fragment), *alr4082*, and *all8559* (see Fig. S7, CL 3); IS1182 family, *alr9024*; ISAs1 family, *all8064*, *all8065* (see the text); IS*Azo13* family, *all2145* (see Fig. S8, CL 2), *asr7385*; IS*H3* family, *all7244*; not classified: *alr7163* (see the text). −, not assigned by ISfinder.

[b] Nunvar et al. (22).

[c] Cluster (CL) identifications are in boldface.

[d] 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; *Cy*7425, *Cyanothece* sp. strain PCC 7425; *Cw*, *Crocosphaera watsonii* WH8501; *Cy*0110, *Cy*7424, and *Cy*8801, *Cyanothece* sp. strains PCC 0110, PCC 7424, and PCC 8801, respectively; *Gvi*, *Gloeobacter violaceus* strain PCC 7421; *Lyn*, *Lyngbya* sp. strain PCC 8106; *Mc*, *Microcoleus chthonoplastes* strain PCC 7420; *Np*, *Nostoc punctiforme* strain PCC 73102; *Ns*, *Nodularia spumigena* CCY9414; *Sy*7002, *Synechococcus* sp. strain PCC 7002; *Te*, *Thermosynechococcus elongatus*. For other ISs, see ISfinder.

[e] Identities to the inverted R end are underlined. Boldface indicates palindromic sequence.

[f] Identities to the L end are underlined. Boldface indicates palindromic sequence.

[g] NSS, not significantly similar.

[h] Confirming Cai (7, 8).

[i] Duplicated sequence is underlined.

[j] A, alignment; K, in known sequence.

[k] aka, also known as.

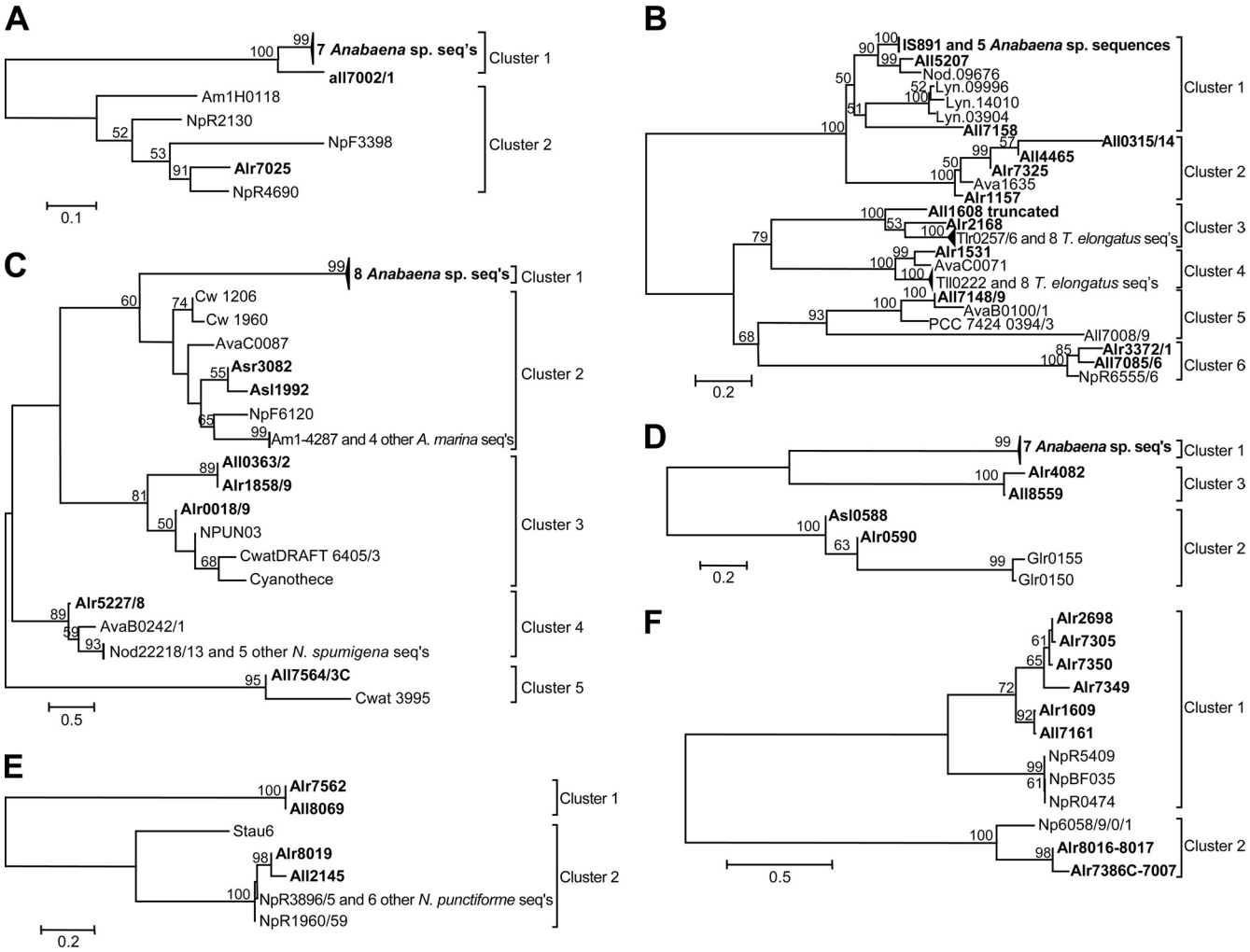


FIG. 1. Phylogenetic relationships of amino acid sequences of transposase genes of *Anabaena* sp. ISs and their homologs in the following families are shown: IS*5* (A), IS*200*-IS*605* (IS*1341*, including IS*891*, and IS*608*) (B), IS*630* (including IS*895*) (C), IS*982* (D), IS*Azo13* (E), and IS*L3* (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.

IS*4* family. *Anabaena* sp. ORFs *alr1332*, *alr5204*, and *all7115* encode transposases in the IS*4* 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

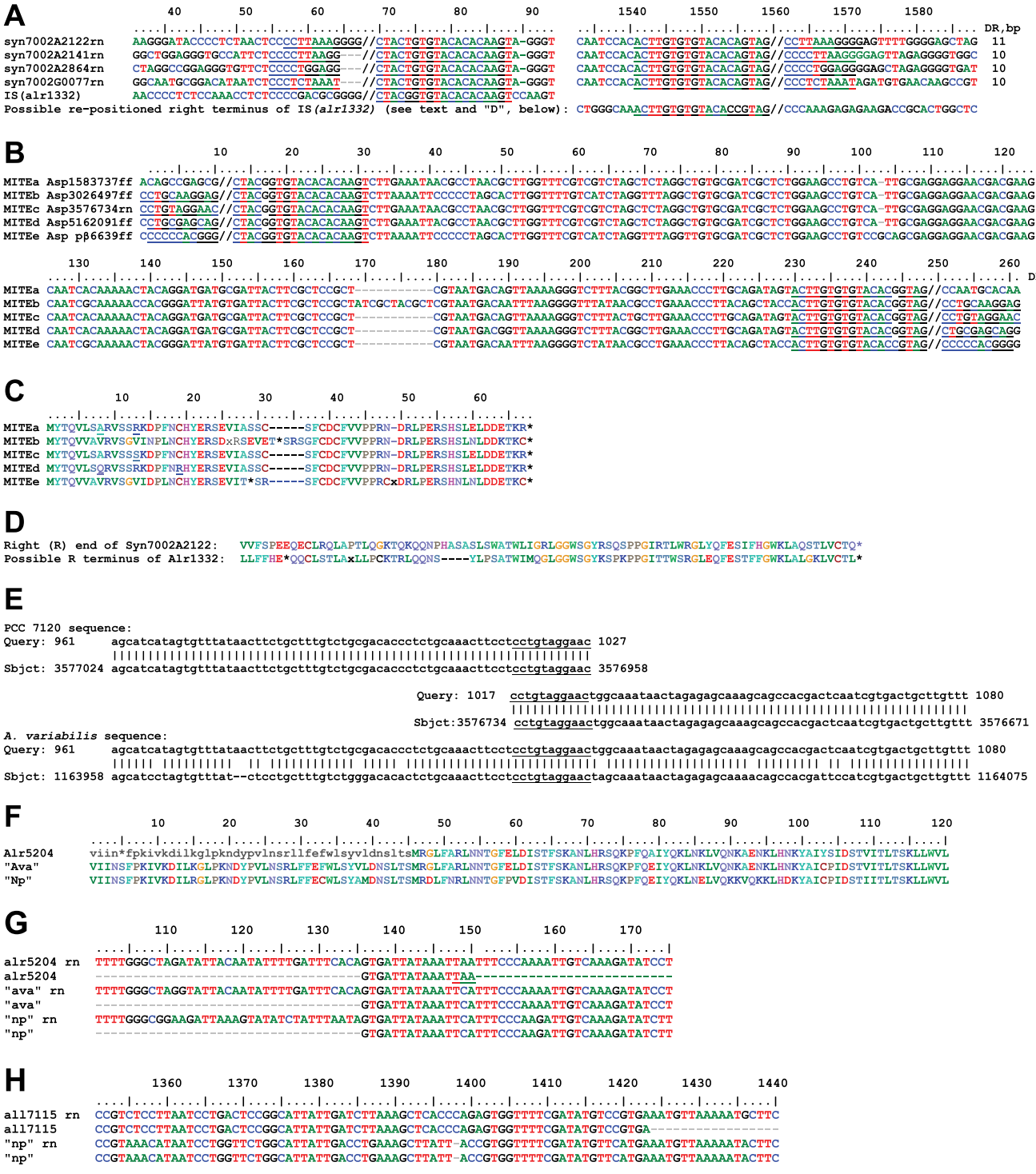


FIG. 2. Transposable elements in the IS*4* family. Border regions of IS(*alr1332*) and related ISs from *Synechococcus* sp. strain PCC 7002 (A) and MITEs whose ends are similar to those of the ISs in panel A are shown (B). //, presumptive ends of ISs and MITEs; rn, region. The Syn7002 ISs extend from 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 *all7115*, 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.

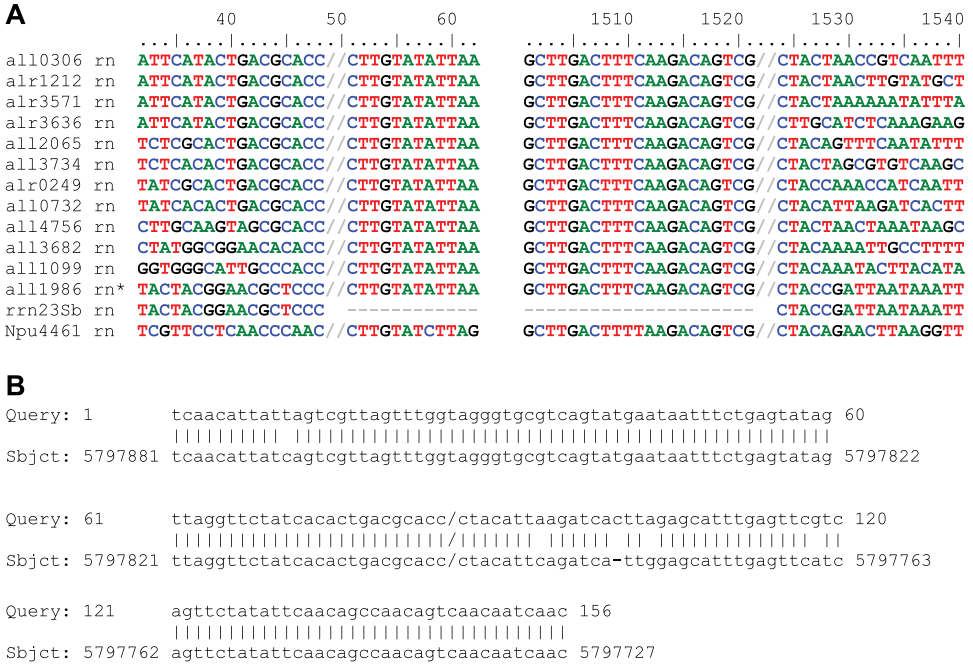


FIG. 3. (A) Border regions of copies of IS*1594.* Presumptive ends (//) of the ISs could be moved to the right 1 bp (C//TTGTAT . . . TCGC//T) or 2 bp (CT//TGTAT . . . TCGCT//). *, IS*1594* copy within *rrn23Sa.* (B) A region of *A. variabilis* (Sbjct, subject) between *ava4633* and *ava4634* has great 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. variabilis*—enabled 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.

IS*5* family, group IS*1031*. Whereas ISs of group IS*1031* 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.

IS*110* family. The only IS*110* family ISs of *Anabaena* sp. are 12 identical copies of IS*1594*, each of which bears 1 of the 30 AvrII sites in its genome (5, 15). The copy reported in GenBank (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 IS*200*/IS*605* family. This family, which includes IS*891* and related ISs (Table 1), normally has one or two ORFs and neither IRs nor DRs (ISfinder). IS*891* (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 IS*891*s. Therefore, these copies of IS*891* may preferentially, perhaps obligatorily, target TTAC sites. With one exception, members of IS*891*-related cluster 2, whose transposases have diverged from those of IS*891* (Fig. 1B), retain TTAC as a presumptive target site, but members of IS*891*-related 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 IS*891* and appears to represent the R end of many other IS*891*-related ISs (Fig. 4A to D and G). Figure 4G presents an apparent exception. Homology of IS(*all7008*, *alr7009*) and other sequences in IS*200*-IS*605* 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 IS*891*-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 IS*Ljo5*, within the IS*200*/IS*605* family. Alr1015 is also related to Alr4734 and has yet greater similarity to transposases of *Cyanothece* spp. and a *Microcoleus* sp. (Fig. 6A). Tyrosine

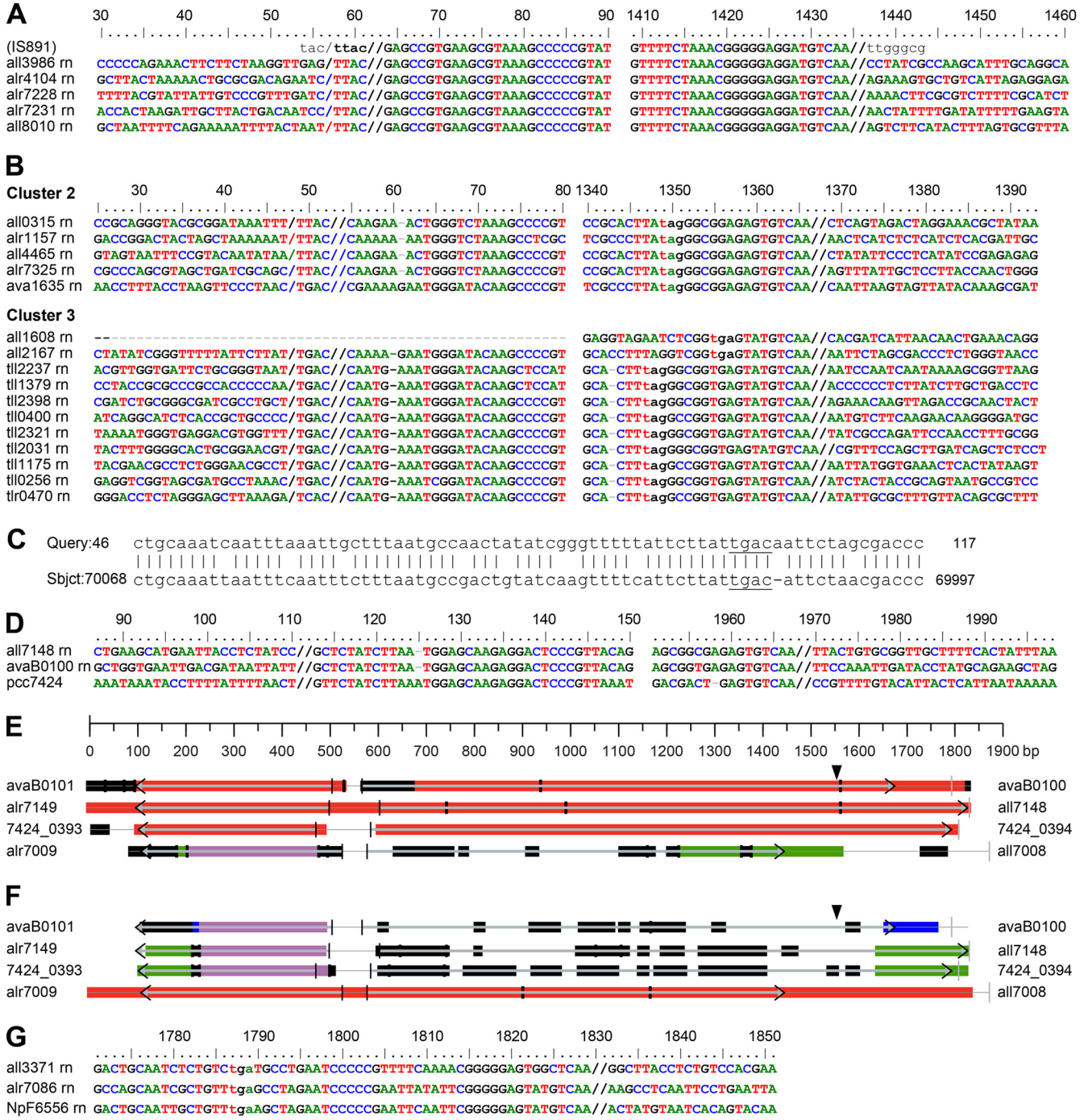


FIG. 4. Characteristics of IS*891*-related ISs. (/, presumptive left end of target site). (A) Border regions of copies of IS*891*. (B) Border regions of members of IS*891*-related clusters 2 (the top five) and 3. Lowercase tag and tga are the termination codons of the right-hand transposase ORF. (C) 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 IS*891*-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, ≥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(alr7386′, 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 IS*200*-IS*605* family shows an unusual example of an IS*891*-related IS in which the R end, near the (lowercase) termination codon tga of *alr3372*, may terminate in a variation of TGTCAA.

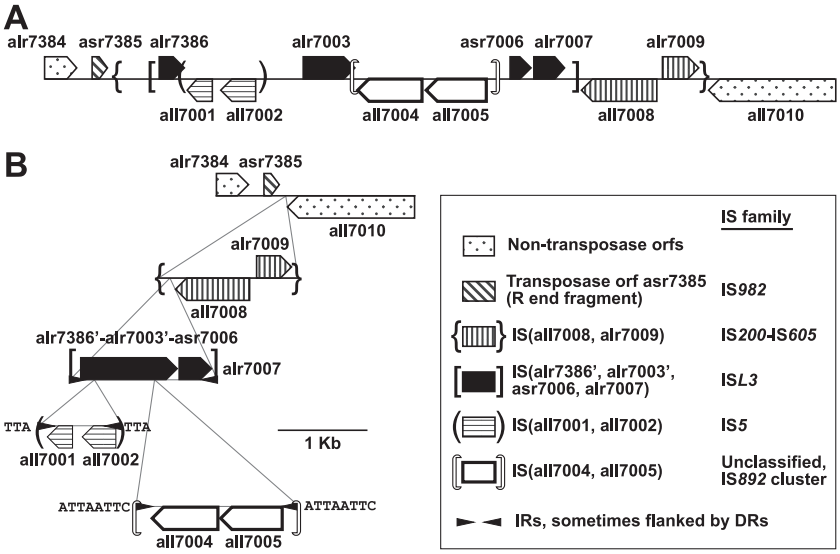


FIG. 5. Stacked ISs. (A) A linear portrayal of a region of plasmid α in the vicinity of its arbitrary origin, a unique SalI site within *all7001*. Except for *alr7007*, ORFs *asr7385* through *alr7009* are annotated as encoding transposases. The ends of the ISs are indicated by individualized brackets. (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(*all7005*, *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 IS*200* (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 (GGT GGGCAATGCCCACC), 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.

IS*630* family. IS*895* and its relatives (1) are part of the IS*630* family, whose members normally have one ORF (ISfinder). However, seven of the eight *Anabaena* sp. IS*895*s 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 IS*895* is thought to incur translational frameshifts (32). A similar indel mutation is seen in Fig. 7B for positions 386 to 393 of IS*891* cluster 2.

IS*630* 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 IS*630*-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 IS*630* members of *Anabaena* sp. (Fig. 8F and Table 1).

The IS*892* cluster of ISs. IS*892* (7, 9) is present in ISfinder within an assemblage of "unclassified IS elements" (17). Whereas copies of IS*1594* and IS*895* are present only in the chromosome of *Anabaena* sp., copies of IS*892* are present only in plasmids α and δ of that strain. The latter localizations and the ability of plasmid α to move from one strain to another (21) suggest strongly that IS*892* reached *Anabaena* sp. within a plasmid. Ten copies of IS*892* 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 IS*895*, those of IS*892* 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

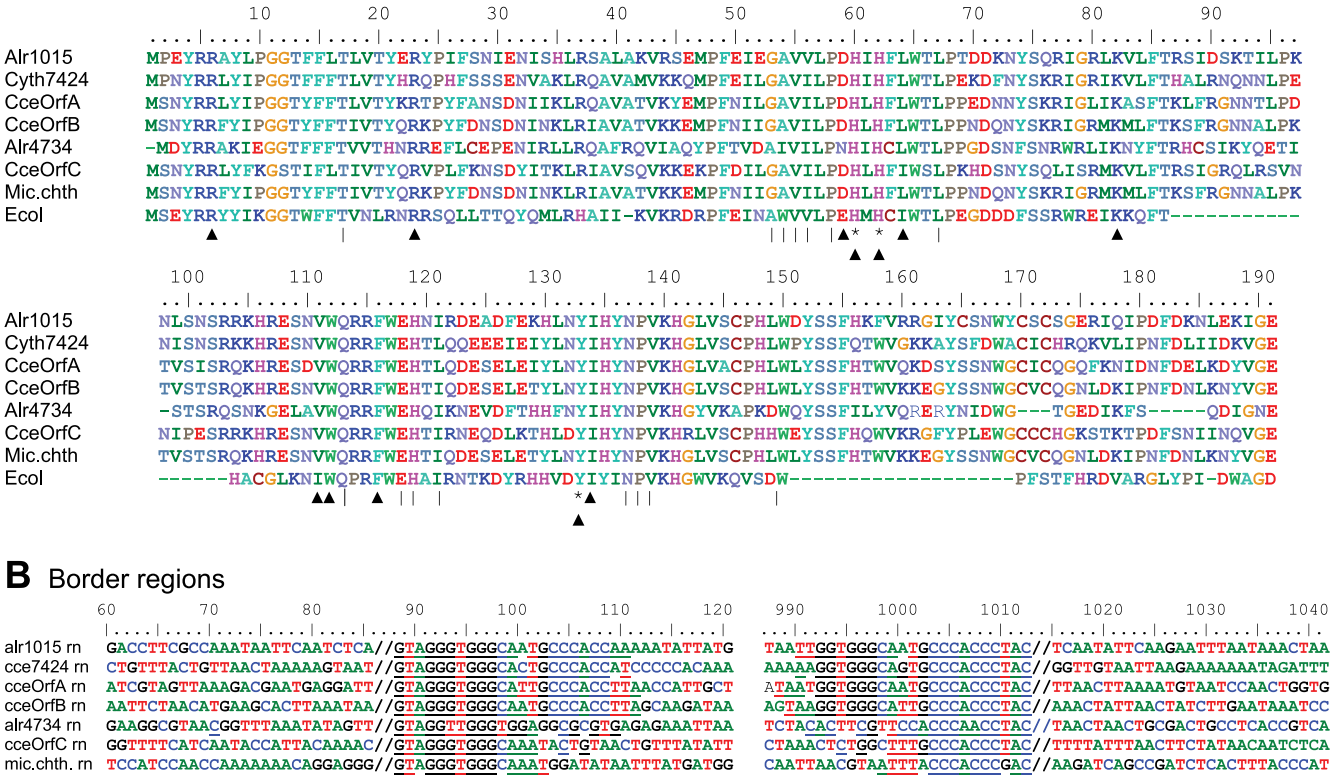


FIG. 6. *Anabaena* sp. RAYTs. Amino acid sequence (A) and nucleotide flanking regions (B) of IS(*alr1015*) and related ISs. Amino acid residues marked with an asterisk (*) are within the catalytic region of the protein, residues marked with a vertical line (|) below the sequence are conserved in tyrosine transposases called RAYTs, and residues marked with black triangles are conserved both in RAYTs and in IS*200* 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.

IS*982* family. Seven IS*982* family members found by ISfinder and denoted IS*Nsp1*s (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*, *all7563′*) (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. punctiforme* 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).

IS*L3* family. Both clusters of *Anabaena* sp. ISs within the IS*L3* 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 IS*L3* 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, IS*493* (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 IS*As1* 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-α 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*).

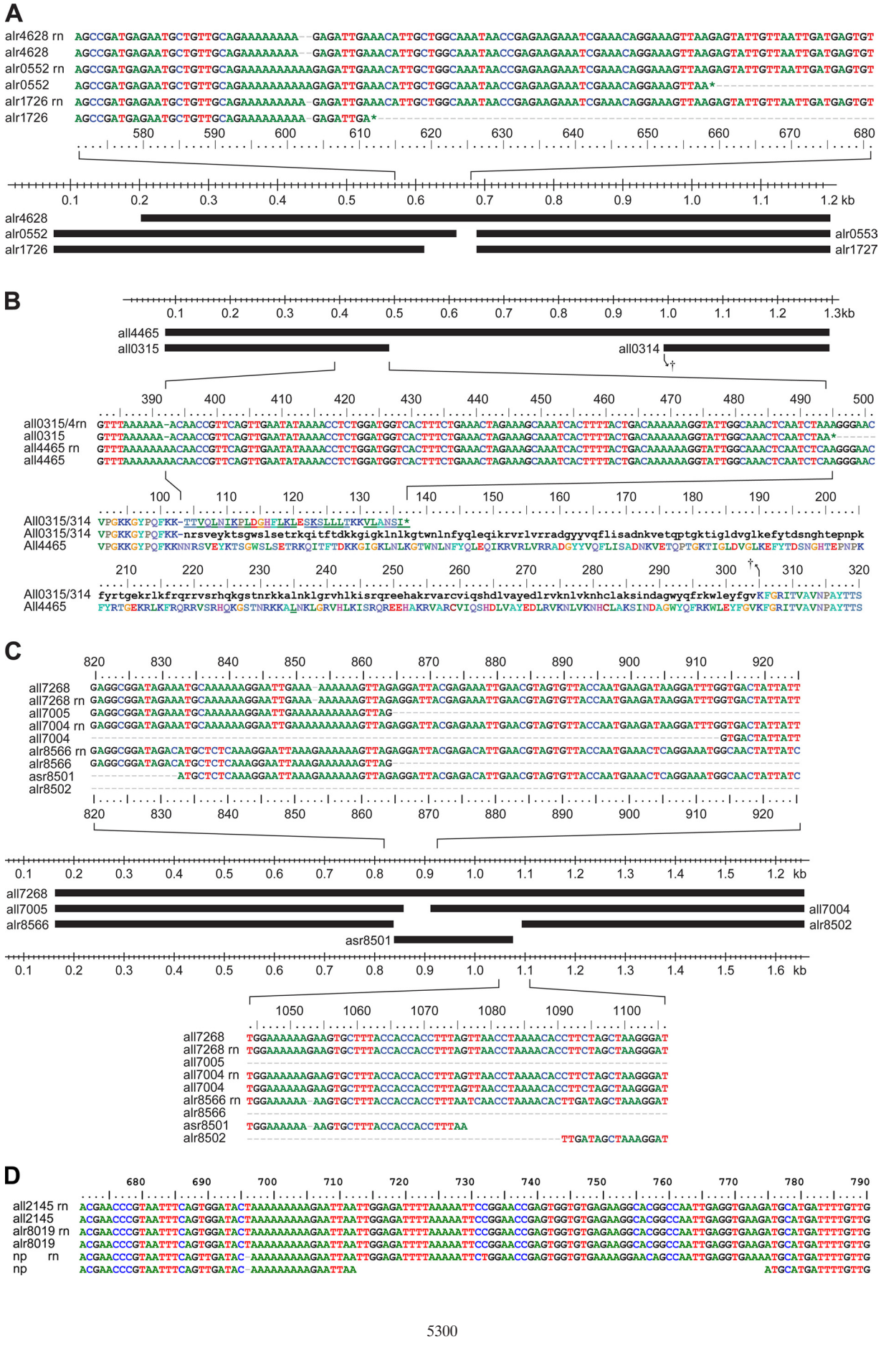
In some instances in which a gene annotated as a "transposase" was near a short (50 to 99 aa) ORF annotated "*asl*…" or "*asr*…" (15), the short ORF comprised a genomic sequence, as it existed prior to insertion of the IS, and a sequence from within an end of the IS that lacked an intervening, in-frame stop codon. IS*1594* thus contributed, in addition to its transposase gene, large parts of ORFs *asl0305* and *asl1098*. More generally, IS*1594* 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. IS*891* 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 IS*891*-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 IS*895* (Fig. 7A), IS*891*-related cluster 2 (Fig. 7B), and IS*892* (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 IS*Azo13* 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.


A
alr4628 rn
alr4628
alr0552 rn
alr0552
alr1726 rn
alr1726
alr4628
alr0552
alr1726
alr0553
alr1727
B
all4465
all0315
all0314
all0315/4rn
all0315
all4465 rn
all4465
All0315/314
All0315/314
All4465
C
all7268
all7268 rn
all7005
all7004 rn
all7004
alr8566 rn
alr8566
asr8501
alr8502
all7004
alr8502
D
all2145 rn
all2145
alr8019 rn
alr8019
np rn
np

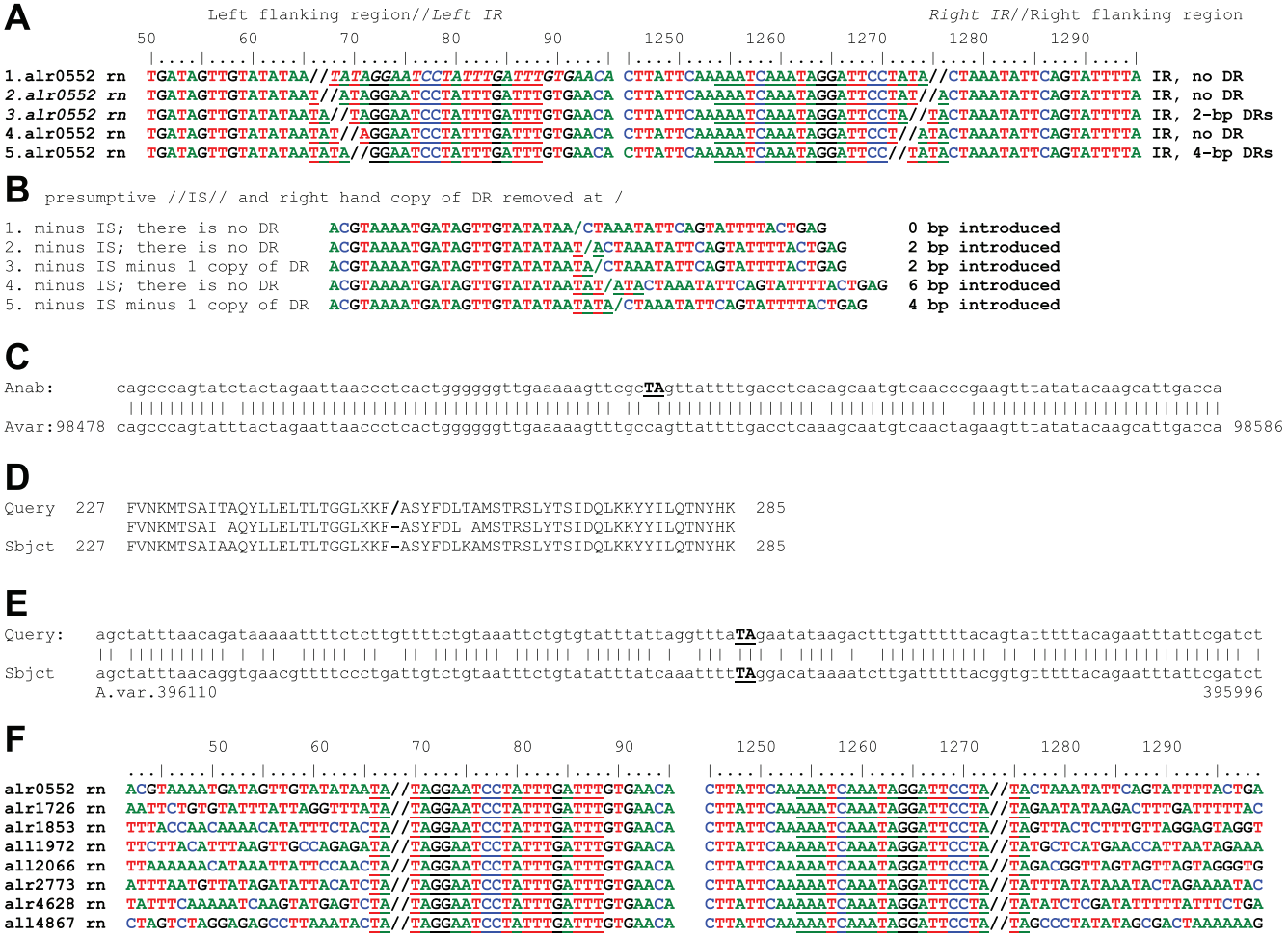


FIG. 8. Palindromes at the ends of IRs of copies of IS*895*-related ISs. (A) If an IS ends in, or is flanked by, a palindrome (here, TA or TATA, and underlined), that palindrome can form part of the IRs (interpretations 1 to 4), DRs (3 or 5), or both (3), resulting, as shown in panel B, in different 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 transposed 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) IS*895*s. 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) IS*891*-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) IS*892*, aka IS(*all7268*), and related ISs. The partial sequence and the bar that represent IS(*all7268*) also 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*, *all7105*), IS(*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 864. 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* IS*s* (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 insertion of ISs

Family/group/CL[a]	Comment
IS*4*/IS*Pepr1*/CL IS*Acma11*	IS(*all7115*) interrupts a GNAT superfamily protein gene, with alignment of homologs Npun_R0240 and PCC7424_2037 across the mutation site.
IS*5*/IS*1031*/CL 1	IS(*alr4438*, *alr4439*) is within a questionably significant, 55-aa ORF. IS(*all4817*, *all4816*) is within a Bpu10I RE Superfamily protein gene (19). IS(*alr5157*, *alr5158*) is within a gene that fuses the start of *alr5156* and the end of *alr5159* and that has homologs in diverse cyanobacteria and other bacteria. BLASTn shows that IS(*all7002*, *all7001*) can be computationally excised from a sequence closely matching that of *alr8016*. Computational removal of all of these ISs leads to no shift in the reading frame.
IS*5*/IS*1031*/CL 2	IS(*alr7025*) revises the N terminus of a membrane protein presumptively encoded by *all7024*.
IS*110*/(−)/IS*1594*	Removing IS(*all0306*) destroys *asl0305*, and removing IS(*all1099*) destroys *asl1098.* IS(*alr3636*) is inserted within an ORF that may encode a DNA mismatch repair protein.
IS*200*-IS*605*/IS*1341*/IS*891*	Each of IS(*all3986*) and IS(*alr4104*) splits a two-component His kinase family gene, and IS(*alr7231*) splits a Ser/Thr kinase family gene.
IS*200*-IS*605*/IS*1341*/IS*891*-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.
IS*200*-IS*605*/IS*1341*/IS*891*-related CL 6	IS(*all7085*, *alr7086*) revises the 5′ terminus of *all7084*.
IS*630*/(−)/IS*895*-related CL 1	IS(*alr1726*, *alr1727)* is within a 102-aa ORF that may not be a gene. IS(*alr1853*, *alr1854*) is within a 68-aa ORF that may encode a protein. Removal of IS(*all1972*, *all1971*) extends the 5′ end of *all1973* to 37 bp from *all1970*.
IS*630*/(−)/IS*895*-related CL 3	IS(*alr1858*, *alr1859*) splits an 88-aa ORF that putatively encodes a protein with a PIN (PilT N terminus) domain in diverse cyanobacteria and other bacteria.
IS*630*/(−)/IS*895*-related CL 5	BLASTn and BLASTp localize IS(*all7564*, *all7563*) in an ORF that encodes an E1 enzyme (super)family protein.
IS*892*	IS(*all7178*, *all7177*) is inserted within an unannotated, 417-bp ORF that overlaps *all7178* and has strong homologs, including a close match within *N. punctiforme* ORF *npR1908*. IS(*all7005*, *all7004*), IS(*all7106*, *all7105*), IS(*all7268*), IS(*all7303*, *all7302*), IS(*alr7323*), and IS(*alr8510*) 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(*all7112*, *all7111*) is inserted 19 bp 5′ from *all7110*, presumably affecting its transcription.
IS982/(−)/CL 1	Removal of IS(*all2664*) led to the finding of a 266-aa ORF (Table S1 in the supplemental material) that fuses the 3′ end of *asr2665* to and through *asr2666*. IS(*all3624*) is positioned within a fusion of *alr3623* and *alr3625* that encodes a type I restriction modification system DNA specificity subunit. Removing IS*Nsp1* (aka IS[*alr1569*]) extends *asr1570* 5′, 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 *ava4176*.
IS982/(−)/CL 3	IS(*alr4082*) sits in an ORF whose predicted product has tBLASTn scores of >6e−25 in other strains.
ISAzo13/(−)/CL IS*Nsp4*	The sequence intercepted by IS(*alr8019*) is highly similar to (89% amino acid identity predicted) but only 63% as long as *all7355*, with alignment across the mutation site. A 426-bp, unannotated ORF that fuses the ends of *alr2144* and *all2145* is highly homologous to *A. variabilis* predicted protein AvaC0094. IS(*all2145*) ends, truncated, 49 bp 3′ from that 426-bp ORF.
ISL3/(−)/CL 1	IS(*alr2698*) is inserted five codons before the end of *alr2697*.

[a] CL, cluster; −, not assigned by ISfinder.

because Alr5204 and All7115 appear to be structurally defective, no IS*4*-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 IS*1594* 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 IS*1594* 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

1. **Alam, J., J. M. Vrba, Y. Cai, J. A. Martin, L. J. Weislo, and S. E. Curtis.** 1991. Characterization of the IS*895* family of insertion sequences from the cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. **173:**5778–5783.
2. **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.
3. **Baltz, R. H., D. R. Hahn, M. A. McHenney, and P. J. Solenberg.** 1992. Transposition of Tn*5096* and related transposons in *Streptomyces* species. Gene **115:**61–65.
4. **Bancroft, I., and C. P. Wolk.** 1989. Characterization of an insertion sequence (IS*891*) of novel structure from the cyanobacterium *Anabaena* sp. strain M-131. J. Bacteriol. **171:**5949–5954.
5. **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.
6. **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.
7. **Cai, Y.** 1991. Characterization of insertion sequence IS*892* and related elements from the cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. **173:**5771–5777.
8. **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.
9. **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.
10. **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.
11. **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.
12. **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.
13. **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.
14. **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.
15. **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.
16. **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.
17. **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.
19. **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.
20. **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.
21. **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.
22. **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.
26. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
27. **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.
28. **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.
29. **Snyder, L., and W. Champness.** 1997. Molecular genetics of bacteria. American Society for Microbiology, Washington, DC.
30. **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.
31. **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.
32. **Zhou, F., V. Olman, and Y. Xu.** 2008. Insertion sequences show diverse recent activities in cyanobacteria and archaea. BMC Genomics **9:**36.
33. **Zhou, F., T. Tran, and Y. Xu.** 2008. Nezha, a novel active miniature inverted-repeat transposable element in cyanobacteria. Biochem. Biophys. Res. Commun. **365:**790–794.
34. **Zhou, R., and C. P. Wolk.** 2002. Identification of an akinete marker gene in *Anabaena variabilis*. J. Bacteriol. **184:**2529–2532.