Very small mobile repeated elements in cyanobacterial genomes

Jeff Elhai, Michiko Kato, Sarah Cousins, et al.

Genome Res. 2008 18: 1484-1499 originally published online July 3, 2008
Access the most recent version at doi:10.1101/gr.074336.107

Letter

Very small mobile repeated elements in cyanobacterial genomes

Jeff Elhai,[1,8] Michiko Kato,[1,2,4] Sarah Cousins,[1,5] Peter Lindblad,[3,6] and José Luis Costa[3,7]

[1]*Center for the Study of Biological Complexity and the Department of Biology, Virginia Commonwealth University, Richmond, Virginia 23284, USA;* [2]*Department of Forensic Science, Virginia Commonwealth University, Richmond, Virginia 23284, USA;* [3]*Department of Physiological Botany, Evolutionary Biology Centre, Uppsala University, Villavägen 6, Sweden*

Mobile DNA elements play a major role in genome plasticity and other evolutionary processes, an insight gained primarily through the study of transposons and retrotransposons (generally ~1000 nt or longer). These elements spawn smaller parasitic versions (generally >100 nt) that propagate through proteins encoded by the full elements. Highly repeated sequences smaller than 100 nt have been described, but they are either nonmobile or their origins are not known. We have surveyed the genome of the multicellular cyanobacterium, *Nostoc punctiforme*, and its relatives for small dispersed repeat (SDR) sequences and have identified eight families in the range of from 21 to 27 nucleotides. Three of the families (SDR4, SDR5, and SDR6), despite little sequence similarity, share a common predicted secondary structure, a conclusion supported by patterns of compensatory mutations. The SDR elements are found in a diverse set of contexts, often embedded within tandemly repeated heptameric sequences or within minitransposons. One element (SDR5) is found exclusively within instances of an octamer, HIP1, that is highly over-represented in the genomes of many cyanobacteria. Two elements (SDR1 and SDR4) often are found within copies of themselves, producing complex nested insertions. An analysis of SDR elements within cyanobacterial genomes indicate that they are essentially confined to a coherent subgroup. The evidence indicates that some of the SDR elements, probably working through RNA intermediates, have been mobile in recent evolutionary time, making them perhaps the smallest known mobile elements.

[Supplemental material is available online at www.genome.org.]

The genomes of eubacteria and archaea are gene dense, with noncoding regions accounting for only 6%–24% of the genome (Mira et al. 2001). The tandem repeats that are frequently seen in eukaryotic noncoding regions (Tóth et al. 2000) have seldom been reported in bacteria (van Belkum et al. 1998), but dispersed repeated sequences are common: transposons and minitransposons (Siguier et al. 2006), long palindromic sequences in the range of from 60 to 200 nt (Wilson and Sharp 2006), short palindromic sequences in the range of from 27 to 60 nt and their composites (Bachellier et al. 1999; Tobes and Ramos 2005), and very short dispersed repeats (Robinson et al. 1995; Smith et al. 1999; Mrázek et al. 2002; Arakawa et al. 2007). Considering the ubiquity of dispersed repeated sequences and their emerging role in genome evolution (Shapiro 2005; Siguier et al. 2006), we can hardly be said to comprehend bacterial genomes without first giving an account of what they are, what they do, and how they arise.

The mechanisms by which new copies of dispersed repeats are generated are well understood in some, but by no means, all cases. Transposon copies arise through DNA intermediates, catalyzed by transposases (Gueguen et al. 2005). Copies of retroposons (and degenerate retroposons, such as *Alu* sequences) are made through RNA intermediates, catalyzed by reverse transcriptase (Ostertag and Kazazian 2001). Minitransposons (also called MITEs), a category that may include the well-studied ERIC sequences (De Gregario et al. 2005), are thought to rely on transposases encoded outside of the mobile element (Siguier et al. 2006). The 8-nt highly iterated palindromic (HIP1) sequences observed in the genomes of many related cyanobacteria are thought not to be mobile but rather to form through mutation from pre-existing sequences (Robinson et al. 1997), The mechanisms by which small dispersed repeats are generated are otherwise unknown.

The genomes of the cyanobacterium *Nostoc punctiforme* ATCC 29,133 and its relatives are unusual in that about 1.5% (~7.5% of intergenic sequences) is taken up by tandem repeats at least 20 nt in length (Meeks et al. 2001; J. Elhai, T. Katayama, R. Narikawa, S. Okamoto, C. Friedland, R. Nayak, M. Ikeuchi, and M. Kanehisa, unpubl.). Repeating units of 7 nt (Mazel et al. 1990; Meeks et al. 2001) are by far the most common, with only a few of the possible families of heptameric repeats accounting for almost all of the occurrences. They, like eukaryotic microsatellites (Tóth et al. 2000), are responsible for hypervariable loci, including the taxonomically useful length polymorphism within the intron interrupting the $tRNA^{Leu}$ (UAA) gene of *Nostoc* strains (Costa et al. 2002). In rare instances, the tandem repeats of the intron are interrupted by a short segment of nonrepeating DNA of 24, 42, 45, or 48 nt in length (Costa et al. 2002). The sporadic

Present addresses: [4]Department of Microbiology, University of California at Davis, Davis, CA 95616, USA; [5]Department of Biology, Virginia Polytechnic University, Blacksburg, VA 24060, USA; [6]Department of Photochemistry and Molecular Science, The Ångström Laboratories, Uppsala University, SE-751 20 Uppsala, Sweden; [7]Institute of Molecular Pathology and Immunology of the University of Porto—IPATIMUP, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal.
[8]Corresponding author.
E-mail elhaij@vcu.edu; fax (804) 828-0503.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.074336.107.

18:1484–1499 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

nature of the interpolations indicates recent origins and raises the question of how such interpolations might arise.

We report here that these short interpolations themselves are dispersed throughout the genome. The analyses of these sequences in context with their surroundings provide insights as to the mechanistic basis of their dispersal, apparently distinct from mechanisms thus far described.

Results

Characterization of iterated sequences in the genome of *Nostoc*

Costa et al. (2002) reported the sequences of the $tRNA^{Leu}$ (UAA) intron from 54 isolates of *Nostoc*. Of these, eight possess segments interrupting a region of tandemly repeated heptamers, with six of the segments being unique. Matches were found to five of the unique segments in the genome of *N. punctiforme,* none in the intron. No match was found to a sixth unique segment, 42 nt in the intron of *Nostoc* strain Nos20, and it was not considered further in this study. The results of these searches are discussed below and summarized in Figure 1 and Supplemental Figures 1–8. Graphical representations of the matches as sequence logos are given in Supplemental Figure 9.

Characterization of SDR1

A loose search of the *N. punctiforme* genome initiated with the identical segments interrupting the $tRNA^{Leu}$ (UAA) introns of *Nostoc* Nos30 and Nos54 brought to light a large and diverse family of 24-nt sequences named SDR1 (for small dispersed repeat; see Discussion). The most commonly encountered family members are shown in Figure 1. The initial search, based on percent identity (see Methods), led to those sequence subfamilies labeled SDR1.1 through SDR1.6, occurring in the *Nostoc* genome as many as 10 times. A more extensive list is provided in Supplemental Figure 1.

The six subfamilies share a conserved core of nine nucleotides or 10 (GAGCG.AG.CGA), if SDR1.3 is excluded. That core is flanked by a 4-nt inverted repeat, again excluding SDR1.3, supported by a pattern of compensatory mutations at positions 5–9 and 21–24. To assess the significance of the inverted repeat, we searched the genome for all instances of the 10-nt core pattern and examined the flanking sequences (Table 1). Far more instances of the core pattern were found than expected by chance, and of the new instances, 77% were flanked by 4-nt inverted repeats of diverse sequences. None would be expected by chance, and we take these results, along with the marked bias toward GT paired substitutions over AC paired substitutions to indicate that selection acts on a single-stranded polynucleotide, presumably RNA transcribed from SDR1, capable of internal base pairing. With SDR1 redefined as possessing the 10-nt core and flanking inverted repeats, we found 11 subfamilies with at least four members (Fig. 1). The definition excludes SDR1.3. This and other special cases are discussed later.

There are rare cases where the 4-nt inverted repeat is greatly extended. In *Crocosphaera watsonii*, six instances of SDR1 have the 4-nt inverted repeat extended nine additional complementary nucleotides in each direction, a conclusion supported by compensatory mutations (Supplemental Fig. 1). Similar extensions can be seen with instances of SDR1 in *Anabaena* PCC 7120 and *Nodularia* CCY9414, but we've seen no such extensions in the many instances of SDR1 in *Nostoc*.

Characterization of SDR2

The identical sequences of *Nostoc* Nos37 and Nos38 that interrupt their $tRNA^{Leu}$ introns were used to search the *Nostoc* genome, yielding a family of sequences, termed SDR2, that fall into two closely related subgroups, SDR2a (24 nt) and SDR2b (25 nt). The two subgroups share the same 20-nt palindromic core, with the exception of an insertion/deletion (Fig. 1). Although the pattern of single mutations within the palindrome is highly biased toward those maintaining base pairing through G-U interaction (A-C on the complementary strand), evidence on the basis of compensatory mutations for a secondary structure of SDR2a and SDR2b maintained by selection is slight (Table 2; Supplemental Figs. 2A, 2B). However, the most common variants of SDR2 differ only in the two central nucleotides (Supplemental Fig. 9C), those that would not be expected to participate in base pairing if the palindrome folded onto itself. The element is sometimes flanked on the right by four nucleotides (often CTAC) capable of extending the palindrome to 28 nt (Supplemental Figs. 2A, 10).

Characterization of SDR3

The interruption in the intron from *Nostoc* Nos51 was used to find in *N. punctiforme* a small family of sequences, SDR3 (Fig. 1; Supplemental Fig. 3). There is insufficient variability in the set to permit identification of conserved complementary regions.

Characterization of SDR4

No sequence in the *N. punctiforme* genome matched the full extent of the 45-nt interruption from *Nostoc* Nos33. An internal 25-nt segment, however, was similar to the largest family of small repeated elements we have found, SDR4 (Fig. 1; Supplemental Fig. 4). The element has two pairs of complementary segments, both confirmed by multiple compensatory mutations that preserve the base pairing (Table 2; Supplemental Fig. 4). Note, however, that the choice of strand shown in Figure 1 is somewhat arbitrary. We cannot exclude the possibility that the complementary strand is functionally important. The remaining part of the sequence from *Nostoc* Nos33 is identical to SDR1.21 (Supplemental Fig. 1). The original 45-nt sequence is therefore a fusion of two SDR elements, a matter that will be explored more fully below.

Characterization of SDR5 through SDR8

In characterizing these families of repeated elements related to known interpolations within cyanobacterial introns, we encountered other juxtaposed repeated elements (Fig. 1; Supplemental Figs. 5–8), which we called SDR5 through SDR8. SDR5, like SDR4, possesses regions confirmed by compensatory mutations (Table 2; Supplemental Fig. 5) that may determine the elements' secondary structures. SDR6 has similar regions, although there is insufficient variability to test the significance of the potential base pairing. These three families of repeated sequences share little common sequence, but their predicted secondary structures—in each case two GyA loops fixed in place by GC-rich stems—show striking similarity (Fig. 2A–C). SDR8 may also possess a GyA loop flanked by a GC-rich stem.

SDR4.12 (Fig. 1) is bigger by 3 nt than conventional SDR sequences by reason of a right loop that does not fit the usual pattern. This and other chance discoveries prompted us to look systematically for instances of SDR4 that had up to three additional nucleotides in the right loop. A total of 33 such instances

were found (Fig. 2E,F; Supplemental Fig. 4), In all cases where one additional nucleotide was found, the loop possessed an extra central pyrimidine. In all cases where three additional nucleotides were found, the extreme nucleotides pair with each other and the internal four follow the same pattern, GTTA, as with one additional nucleotide. The left loop seldom exceeds 3 nt in in-

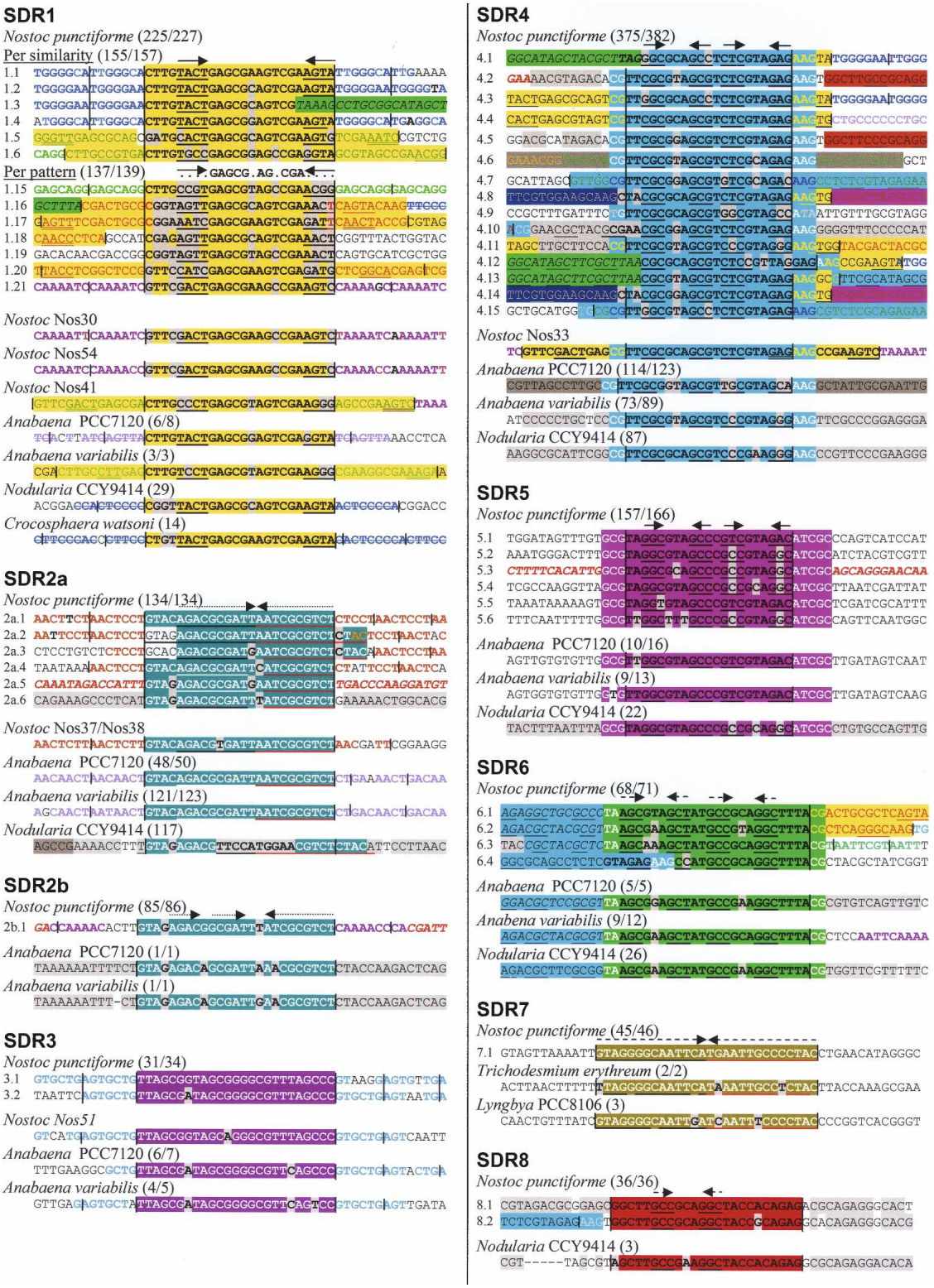


Figure 1. (Legend on next page)

stances of SDR4 in *N. punctiforme*. However, such extensions are much more common in SDR4 sequences in other related cyanobacteria, with their sequences similar to those of right loops (Fig. 2D). We could find no instances of SDR5 with loop extensions.

SDR7, like SDR2a, contains a long perfect palindromic sequence, the self-pairing of which is not significantly supported by an analysis of mutations within the palindrome (Table 2; Supplemental Fig. 7).

Possible secondary structure in SDR elements

All of the SDR families possess palindromic regions capable of internal base pairing, but DNA palindromes are also commonly used as binding sites by dimeric proteins. To assess whether these regions play a role in secondary structure, we analyzed the occurrences of compensatory mutations, important indicators of nucleotide–nucleotide interactions (Rousset et al. 1991; Rivas and Eddy 2001). The number of compensatory mutations in nucleotide pairs postulated to be important in structure (indicated in Fig. 1) were compared with the number from all other possible pairings (Table 2). In the cases of SDR1, SDR4, and SDR5, the number of compensatory mutations at the structural positions significantly exceeded that expected from arbitrary pairings, and in the case of SDR2a, the excess was weakly significant.

Base pairing between G and T (or U in RNA) may also serve to preserve secondary structure or may be an intermediate state on the way to full compensatory mutations (Rousset et al. 1991). A change producing an AC combination is consistent with GT pairing of nucleotides on the opposite strand. Interestingly, the number of single mutations allowing GT pairing in structural positions significantly exceeded expectation in the cases of SDR1 and SDR5, while the number of GT pairings in structural positions on both strands significantly exceeded expectation in the case of the symmetrical element SDR2a. Curiously, the asymmetrical element SDR4 also had significantly increased possibility of GT pairings on both strands at structural positions.

Exhaustive search for SDR elements in the genome of *Nostoc*

We were concerned that by following the interpolations in the *Nostoc* introns, we may have ignored repeated sequences of greater quantitative importance. To address this issue, we performed an unbiased, exhaustive search for elements of at least 24 nt that occur in multiple copies in the genome of *N. punctiforme*, and for comparison, applied the same analysis to genomes from two other bacteria: a distantly related cyanobacterium, *Synechocystis* PCC 6803, and *Escherichia coli* (Table 3).

Topping the list of repeated sequences in *N. punctiforme* are two 24-mers contained within multiple CRISPR families previously described (Godde and Bickerton 2006) in the *Nostoc* genome.

Table 1. Palindromic pairs within instances of SDR1 found by pattern

		Relationship of positions 5–8 and 21–24			
	Total	Paired[a] (exact)	Paired[a] (GT OK)	Paired[a] (AC OK)	Not paired[a]
All instances of SDR1 per pattern					
Total	162	125	22	2	13
Expected[b]	8	0	0	0	8
Found previously[c]	79	61	18	0	0
Not found previously[c]	83	64	4	2	13
Instances of SDR1 per pattern, with copy number >1					
Total	111	96	15	0	0
Expected[b]	0	0	0	0	0
Found previously[c]	66	51	15	0	0
Not found previously[c]	45	45	0	0	0

[a]A sequence found by the pattern . . . . (. . . .)GCGAG.AG.CGA(. . . .), where a period is satisfied by any nucleotide, is counted as "paired" if the two tetranucleotides within the parentheses can base pair with each other either using conventional base pairing or allowing GT or AC pairing. No instance with GT or AC pairing had more than one such pair.
[b]The number expected is: *length* (s^7 w^3), where *length* is the length of the *N. punctiforme* genome (9.1 Mbp), *s* is the fraction of G and the fraction of C in the genome (0.21), and *w* is the fraction of A and the fraction of T (0.29). The probability that a random tetramer pairs exactly with another random tetramer is 0.4%. The same probability allowing for GT pairing at any position or AT pairi006Eg at any position is 2.0%.
[c]SDR1 instances found by the pattern were divided into two groups: (1) those that had previously been found by overall similarity, as described in the Methods section, and (2) those that had not been found by this means.

Tandem heptameric and octomeric repeats and insertion sequences (full-size and miniature) are also on the list, but the majority of common repeated sequences in *Nostoc* are part of SDR elements characterized in the previous section. The lone exception is part of a long imperfect palindrome of variable length, termed SDRQ, that is reminiscent of BoxC sequences from *E. coli* (Bachellier et al. 1999), in that the palindrome arises from a purine-rich region juxtaposed to a pyrimidine-rich region. There are at least 14 copies of this element >100 nt and many dozens of other instances that are smaller or too divergent to be detected by BLAST to their full extent. A more complete characterization of this element (and others of lower copy) lies outside the scope of this report.

Synechocystis PCC 6803 and *E. coli* have quite different patterns of repeated sequences (Table 3). The set of repeated sequences from *Synechocystis* is dominated by insertion sequences, while almost all repeated sequences in *E. coli* with copy number >10 fall into a single previously described family, PU/REP (Bach-

Figure 1. Common instances of SDR elements. Families: The most representative examples are shown of members of different families of SDR elements and their subfamilies. All families with at least four occurrences in the genome of *Nostoc punctiforme* are given, along with examples from every cyanobacterial genome considered (listed in Fig. 5) in which an instance of the SDR element was observed (see Supplemental Files 1–8 for more complete listings). The name of each organism is followed by a fraction consisting of the number of instances of the SDR element in the chromosome divided by the total number of instances. Only the total is given for organisms whose genome sequences have not been completely determined. SDR elements identified in the survey of the $tRNA^{Leu}$ (UAA) intron of different *Nostoc* strains (Costa et al. 2002) are also shown. Color conventions: Each SDR element is highlighted with a characteristic color, used in other figures as well. Deviations from the reference sequence of the group are shown by gray highlighting. Nearby SDR elements are also highlighted in their characteristic color. For these elements, a green or black font is used if the orientation of the element is the same as the central element, otherwise a red font or italics are used. Sequences that are italicized and in red font are parts of putative minitransposons. Blocks outside of the central element highlighted in gray are unnamed sequences found multiple times in the genome. Nearby tandemly repeated sequences are colored and in bold, and the repeating units are separated by vertical bars. Vertical bars also mark the end of SDR elements. Tandem repeats with lines running through them are the inversion of repeats represented by the same color. Inverted repeats: Underscores indicate regions potentially involved in self-base pairing, and extended arrows at the *top* of the grouping indicate whether the pairing is strongly supported by compensatory mutation (solid arrows) or not (broken arrows).

Table 2. Analysis of mutations within regions of putative secondary structure

		Double mutations[b]				Single mutations[b]					
		Total mutated pairs[c]		Compensatory[d]		Total mutated pairs[c]		→ GT[d]		→ AC[d]	
Family	Motifs[a]	Structural	Nonstructural	Actual	Expected	Structural	Nonstructural	Actual	Expected	Actual	Expected
SDR1	52	30	236	**27****	10.3	26	1096	**12***	6.4	7	7.7
SDR2a	83	8	97	**6***	2.9	97	1193	**37****	16.9	**33****	16.8
SDR2b	29	2	6	1	0.7	30	278	8	7.1	14	12.0
SDR3	20	0	8	0	0.0	11	227	5	3.7	4	4.6
SDR4	194	250	894	**217****	69.1	122	4501	**50****	19.1	**28****	17.3
SDR5	96	58	116	**57****	31.5	67	1182	**36****	23.5	19**↓	29.6
SDR6	23	0	0	0	0.0	18	245	3	4.6	6	3.5
SDR7	46	8	33	6	5.3	78	498	34	31.3	27	25.5
SDR8	36	0	8	0	0.0	10	270	2	2.4	6	4.3

[a]One sequence from each subfamily was considered, excluding those sequences with insertions or deletions. In particular, elements with known insertions of SDR elements (e.g., SDR1.3) were excluded. In the case of SDR1, only those instances that were detected according to the number of mismatches were considered.
[b]All pairs of positions in the set of aligned sequences were considered, whenever one nucleotide of the pair was a G and the other a C or one was an A and the other a T. Double mutations were defined as those in which both nucleotides of a pair under consideration differed from the nucleotides at the same positions in the canonical sequence for the SDR family. Single mutations were defined as those in which only one of the two nucleotides differed from the canonical sequence.
[c]Pairs of nucleotides were considered structural if they corresponded to a pair identified in Figure 1 as possibly participating in base-pairing. Otherwise, the pair was considered nonstructural.
[d]Double mutations were considered compensatory if the resulting pair consisted of a G and a C or an A and a T. Pairs suffering single mutations were counted as either leading to a GT pair, an AC pair, or all other pairs. The actual count is the number of instances of the type of mutation under consideration that were observed in structural pairs. The expected count is the number calculated for structural pairs using the mutation-class frequencies in nonstructural pairs. A χ^2 test was used to assess significance. One and two asterisks indicate significance to the $P < 0.05$ and $P < 0.01$ levels, respectively. Significant differences are all increases (shown in bold) except where indicated.

ellier et al. 1999). Members of this family differ from one another only by a few nucleotides, and like several of the SDR elements, PU/REP elements have sequences that assume secondary structure, as confirmed by compensatory mutations (Supplemental Fig. 11).

Since SDR2a and SDR7 contain long perfect palindromes, we also searched the genome for all perfect palindromic sequences of length equal to or greater than 2 × 10 nt (Supplemental Fig. 10). Of the 124 palindromes found, 55 are accounted for by SDR2a and SDR7 sequences, and none of the remainder occur with a copy number greater than 2.

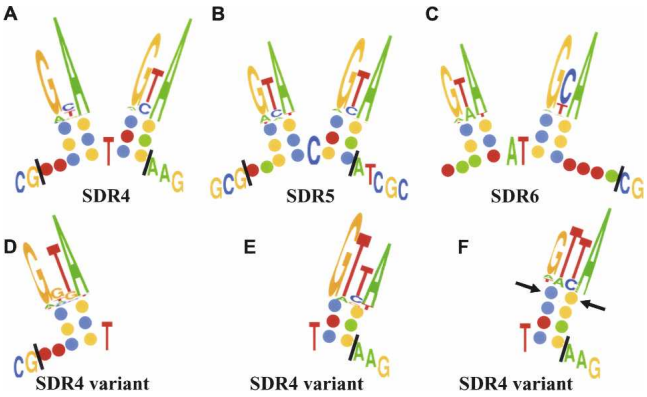


Figure 2. Proposed secondary structures of SDR elements. Loops of elements are given as sequence logos, where the height of the stacked letters at a given position is proportional to the information content at that position and the relative frequency of each nucleotide is given by the relative height of the letter. Consensus nucleotides are represented as either letters or filled circles (A, green; C, blue; G, yellow; T, red). Proposed target sequences are separated from the SDR elements by thick bars. (*A–C*) Complete SDR elements SDR4, SDR5, and SDR6; (*D*) variant left loop of SDR4; (*E,F*) variant right loop of SDR4. The arrows identify an extension of the right stem.

The SDR sequences in Figure 1 are therefore the most numerous of all uncharacterized repeats and palindromes in the genome of *N. punctiforme*, but this type of repeat is not generally found in bacterial genomes. The total number of repeats denoted in Figure 1 represents about 0.3% of the total genome of *N. punctiforme* and about 1.5% of its intergenic sequence.

Locations of SDR elements

It would be of interest to know how rapidly in evolutionary time new insertions occur in the genome. One useful time marker is the typical transit rate for plasmids. If insertion occurs rapidly with respect to the rate of plasmid gain and loss, then one would expect to find SDR elements appearing in the chromosome and in plasmids to a degree proportional to their target sizes. The ratio of the chromosomal target size to plasmid target size may be estimated as the ratio of their lengths, which is 10:1 in the case of *N. punctiforme*. This is probably an overestimate, as plasmids figure to have a higher percentage of genes not subject to selection within *Nostoc*. Insertions of SDR5 are distributed amongst the chromosomes and plasmids, with 5% in the latter (nine out of 166; Fig. 1). In the related cyanobacteria *Anabaena* PCC 7120 and *Anabaena variabilis*, six of 16 insertions of SDR5 are in plasmids, five of them in a single plasmid, pAlpha. The sequences of the remaining SDR families, however, are overwhelmingly in the chromosome, 1001 instances compared with 17 in plasmids instances (a 60:1 ratio). If SDR elements can move between distinct molecules, they move considerably more slowly than the rate of plasmid loss and acquisition, except for SDR5.

As one would expect, there is a strong bias against the insertion of SDR elements within genes (Table 4). This bias is most pronounced when the length of the element is not a multiple of three (Table 4, cf. the fraction of SDR1, SDR2a, SDR3, SDR4, and SDR5 insertions in genes with the same fraction of SDR2b and SDR7 insertions).

Table 3. Most numerous 24-nt sequences in *N. punctiforme*

Rank	Copy	Type of Repeat[a]	Unit Length[b]	Coordinates (one example)[c]
Nostoc punctiforme (9,059,191 nt)				
1	84	CRISPR portion	37 nt	1229318..1229341
2	50	**SDR2b**	25 nt	4617288..4617312
3	43	CRISPR portion	37 nt	3341103..3341126
4	37	**SDR5.1**	21 + 8 nt	3322670..3322693
5	36	$(AATTCGT)_n$	7 nt	8232403..8232449
6	30	**SDR6**	28 nt	4586640..4586667
7	29	IS-Np1A1	1005 nt	7588577..7589581
8	29	IS-Np2A1	1265 nt	3873350..3874614
9	20	**SDR2a**	24 nt	6741441..6741465
10	19	$(AGCAGGGG)_n$	8 nt	3100463..3100544
11	19	$(AATGACT)_n$	7 nt	412848..412919
12	19	SDR Q	<147 nt	7141887..7142017
13	19	**SDR4.1**	21 + 5 nt	6894594..6894619
14	17	$(AATGGGG)_n$	7 nt	1042689..1042741
15	16	**SDR4+**	21 + 5 nt	965504...965527
16	15	IS-Np10A-m	120 nt–154 nt	7258495..7258666
17	15	**SDR4.2**	21 + 5 nt	548981..549004
18	15	**SDR4.3**	21 + 5 nt	4987480..4987503
19	15	Probable mini-IS	92 nt–120 nt	2187313..2187429
20	14	**SDR8**	24 nt	8006879..8006902
Synechocystis PCC6803 (3,956,956 nt)				
1	56	CRISPR portion	37 nt	pSYSA:68,572..68608
2	50	CRISPR portion	37 nt	pSYSA:19366..19402
3	39	CRISPR portion	36 nt	pSYSA:91116..91151
4	24	ISY100	945 nt	1626091..1627040
5	16	ISY523	870 nt	3096318..3097194
6	14	Probable mini-IS	72 nt–183 nt	17686..17812
7	12	ISY203	1173 nt	2326926..2328099
8	11	Probable mini-IS	72 nt–183 nt	114689..114815
Escherichia coli (4,639,221 nt)				
1	75	PU-Y(G)	35 nt	1814208..1814242
2	55	PU-Y(A)	35 nt	2536556..2536590
3	47	PU-Z1	31 nt	2652991..2653020
4	28	PU-Z2(G)	37 nt	2302492..2302528
5	17	CRISPR	27 nt	2902036..2902428
6	12	PU-X	35 nt	4151435..4151469
7	12	PU-Z2(GA)	37 nt	2116585..2116621
8	12	PU-Z2(T)	37 nt	4247384..4247420
9	11	IS5	1195 nt	1394068..1395262

[a]SDR elements described in this work are shown in bold.
[b]The length given for SDR elements is the length of the element plus its proposed target sequence.
[c]The coordinates are chromosomal unless otherwise indicated.

The location of the palindromic elements SDR2a, SDR2b, and SDR7 appears to be biased toward sequences downstream from two genes (Table 4). Part of the bias may be an artifact of the greater percentage of space within convergent genes that is not subject to strong selective pressure, but if so, then the bias is much less with the nonpalindromic SDR elements. Such a bias is seen with palindromic sequences from a wide variety of eubacteria (Petrillo et al. 2006), and in particular, with the PU/REP elements of *E. coli* (Table 4). SDR2a elements go still further, tending to lie close to the ends of genes. A total of 31% are found within 30 nt from the 3′ end of a gene, two to three times the typical value for SDR elements. This characteristic is accounted for by a subset of SDR2a elements, those flanked by heptameric repeats, as described in the next section.

Rho-independent transcriptional terminator sequences typically consist of a long palindrome downstream from a gene, followed by a string of thymidine residues (Yarnell and Roberts 1999). Long thymidine-rich regions (as previously defined by de Hoon et al. 2005) are seldom observed adjacent to the palindromic SDR elements (Supplemental Figs. 2A, 2B, 7) or PU/REP elements (Supplemental Fig. 11), but short stretches (at least four thymidines) are common 3′ to SDR7 sequences, found within 10 nt of 39% of SDR7 sequences, as compared with 3% expected from mononucleotide frequencies. Curiously, in 82% of these cases, the thymidines are immediately preceded by one of the three termination codons, even though the codon is not part of an open reading frame.

Sequences flanking SDR elements

Do the SDR elements as defined constitute mobile units? If an element were dispersed as an independent unit, then one would not expect to see any correlation between variations in the flanking sequences and minor variations within the intervening sequence itself (except as determined by target preferences). On the other hand, if the intervening sequence were dispersed in the genome through a mechanism that depended on flanking sequences, then grouping the SDR elements by their internal variation should simultaneously group the segments by their flanking sequences.

In some cases, there is a strong correlation between an SDR family and a specific flanking repetitive sequence. A total of 68%

Table 4. Context of repeated elements in genome of *N. punctiforme* and *E. coli*

Family	Length[a]	Total[b]	Copy[b] ≥ 3	In genes[c]	Intergenic[c] P	Intergenic[c] C	Intergenic[c] D
Nostoc punctiforme							
SDR1	24 + 0	227	104	30%$^{**\downarrow}$	48% (68%)	10% (14%)	13% (18%$^{**\downarrow}$)
SDR2a	24 + 0	135	54	20%$^{**\downarrow}$	51% (64%)	19% (24%$^{**\uparrow}$)	10% (12%$^{**\downarrow}$)
SDR2b	25 + 0	86	54	1%$^{**\downarrow}$	52% (53%)	41% (41%$^{**\uparrow}$)	6% (6%$^{**\downarrow}$)
SDR3	(24 + 1)+0	34	13	18%$^{**\downarrow}$	50% (61%)	15% (18%)	18% (21%)
SDR4	21 + 5	382	171	14%$^{**\downarrow}$	54% (63%)	14% (17%)	17% (20%$^{**\downarrow}$)
SDR5	21 + 8	166	70	36%$^{**\downarrow}$	30% (46%)	14% (23%$^{*\uparrow}$)	20% (31%)
SDR6	25 + 4?	71	45	3%$^{**\downarrow}$	69% (71%)	14% (14%)	14% (14%$^{*\downarrow}$)
SDR7	~28	46	9	0%$^{**\downarrow}$	48% (48%)	39% (39%$^{**\uparrow}$)	13% (13%)
SDR8	24|57	36	18	3%$^{**\downarrow}$	75% (77%)	14% (14%)	8% (9%)
Total		1185	538	19%$^{**\downarrow}$	50% (61%)	17% (21%)	14% (18%)
Fraction of genome in given context				81%	11% (57%)	3% (13%)	6% (29%)
Escherichia coli							
PU[d]	30–37	241	184	7%$^{**\downarrow}$	45% (49%$^{*\downarrow}$)	47% (50%$^{**\uparrow}$)	0% (0%$^{**\downarrow}$)
Fraction of genome in given context				89%	6% (55%)	1% (12%)	4% (33%)

[a]The length is given as the number of nucleotides inserted within a target of known sequence (either a conserved gene or a repeated sequence). In the case of SDR3, 25 nt are apparently inserted with the loss of 1 nt from the target. The length of SDR7 may vary. The 24-nt sequence given for SDR8 may be part of a larger element, as discussed in the text. The number following the plus sign (when present) indicates the length of the preferred target. A question mark indicates that the value is based on limited evidence.
[b]Total number of instances (as defined in Methods) and total number where the copy number of identical sequences is at least 3, as taken from Supplemental Tables 1–8.
[c]Context of insertions: at least partly within a gene or intergenic. If intergenic, then between genes in parallel (P), convergent (C), or divergent (D) orientation. The first percentage is relative to all insertions of the given element, and the second is relative to those insertions in intergenic regions. One and two asterisks indicate significance as determined by a χ^2 test at the $P < 0.01$ or $P < 0.001$ level, respectively. The arrow indicates the direction of deviation from expectation.
[d]Only those PU elements containing high-copy 24-mers (Table 4; Supplemental Table 10) found by the search described in the text are considered here.

of those instances of SDR2a that occur in high copy (three or more copies) are flanked by a recognizable variation of STRR8 (Meeks et al. 2001), the heptameric repeat $[AACTCCT]_n$ (34 of 50 cases) (Fig. 1; Supplemental Fig. 2A). A total of 92% of those instances of high-copy SDR3 are flanked by a variant of $[AGTGC TG]_n$ (12 of 13 cases). In other cases, a flanking heptameric repeat is associated with one subfamily and a different repeat with a different subfamily, for example, in the cases of SDR1.2 and SDR1.3 associated with STRR1 (Mazel et al. 1990; Meeks et al. 2001), $[TGGGGAA]_n$, (11 of 15 cases), and SDR1.15 associated with $[GAGCAGG]_n$ (10 of 10 cases).

Some SDR elements are embedded in larger entities that bear the hallmarks of transposable elements (Fig. 3). SDR2b often lies within a larger relatively conserved sequence, typically 101 nucleotides in length, but as many as 134 nucleotides (Supplemental Fig. 2B). The larger units contain a second, inverted copy of SDR2b. The units are bound by constant 18-nt inverted repeats and flanked by a 10-nt direct repeat whose sequence varies from instance to instance. At least one recognizable end can be found next to 72 of the 84 instances of SDR2b. Since inverted repeats flanked by direct repeats are a common signature of transposons, we examined the genome for other instances of the inverted repeat sequences. We found fourfold more units with these inverted ends beside those containing SDR2b. A total of 90% of these were smaller than 160 nt, but two were full-sized insertion sequences containing a copy of the transposase encoded by *npr1652*. In addition, there are 14 full instances of SDR8 and 14 full instances of SDR4 within *npr1652*-related minitransposons (Fig. 3A; Supplemental Fig. 8). Evidently, a minitransposon related to the one containing *npr1652* houses many SDR elements. The minitransposon has been recently reported in *Nostoc* and other bacteria (Zhou et al. 2008).

SDR4, and especially SDR8, are common inhabitants of other minitransposons. Eleven full instances SDR8 and six full instances of SDR4 (all adjacent to SDR8) are found within a minitransposon for which a full transposon is not known, and five more instances of SDR8 are found within a third apparent minitransposon (Fig. 3B; Supplemental Fig. 8). These two minitransposons are also found containing SDR8 within another heterocystous cyanobacterium, *Nodularia* CCY9414. Strikingly, all instances of SDR8 and some of SDR4 in minitransposons lie embedded within the same 33-nt sequence. This sequence occurs uninterrupted eight times in the *Nostoc* genome: three times within the *npr1652*-related minitransposon and five times elsewhere in the chromosome. In all cases, it is flanked on the 3′ end by a purine-rich region, just as it is in the three minitransposons. Conceivably, this region was instrumental in the events leading to the insertion of the SDR elements into the minitransposons. Many instances of the minitransposon shown in Figure 3B are found in the two *Anabaena* genomes, but neither the 33-nt conserved sequence nor SDR8 is found within them (data not shown).

Only four instances of SDR2a in *N. punctiforme* lie within the complete minitransposon shown in Figure 3C, but related cyanobacteria have many more copies (as described later), including a copy of an apparently full transposon. Five instances of SDR5.3 lie within full-length transposons (Fig. 1; data not shown).

The sequences surrounding SDR2a elements are strongly associated with their positional characteristics. The bias of SDR2a elements for positions immediately downstream from genes described in the previous section is accounted for completely by the subset that are surrounded by heptameric repeats. In this subclass, 57% of the SDR2a elements are within 30 nt of the 3′ end of a gene, and most remarkably, in all but one of the 40 instances, the gene is situated to the right, according to the orientation shown in Figure 1 and Figure 6C (below) (see Supplemental Fig. 2A for multicopy instances).

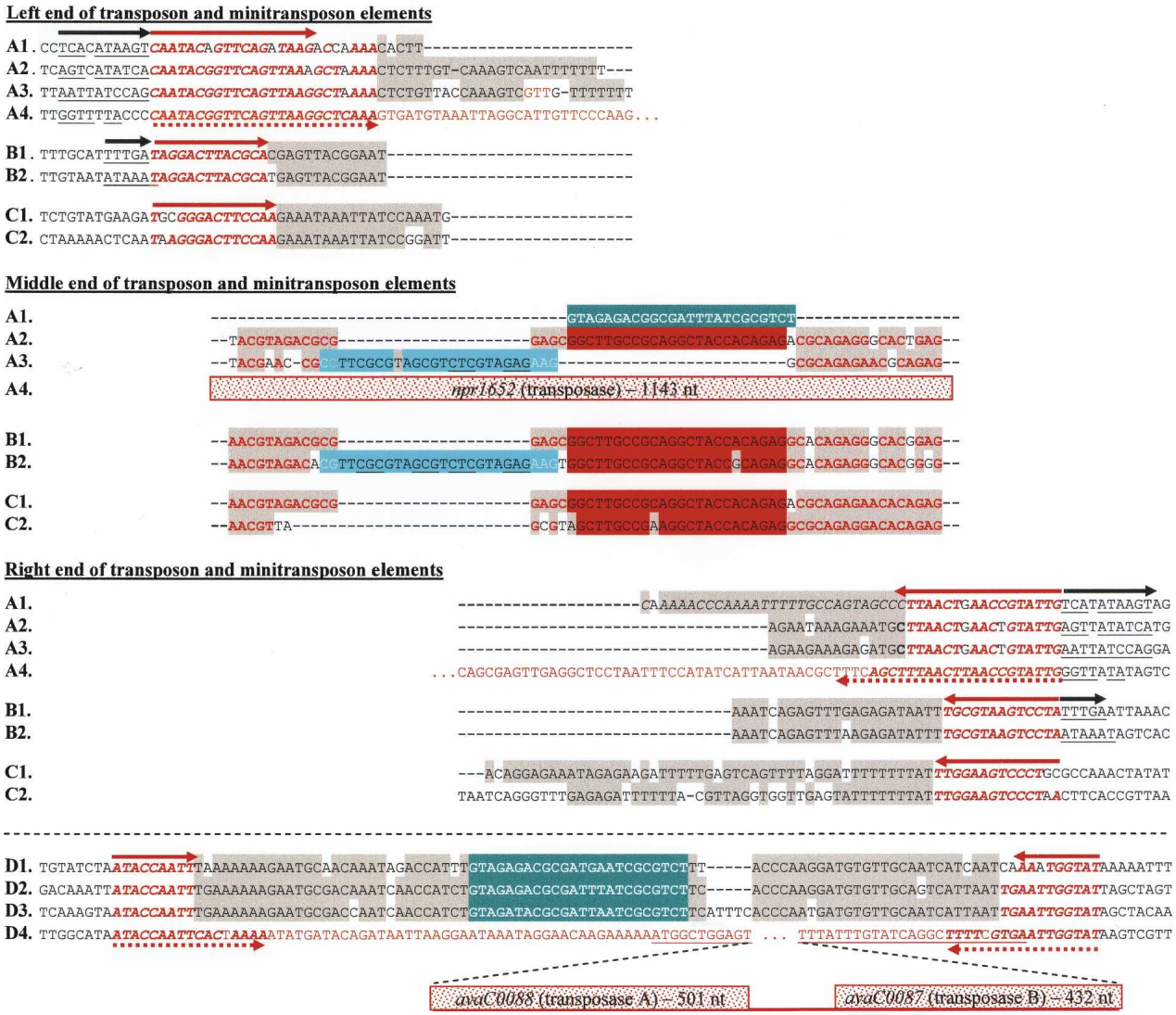


Figure 3. Minitransposons carrying SDR elements. Representative examples of four families (A–D) of minitransposons are shown. When a parent full transposon has been identified, its sequence is in red as well as identical sequences in the derivative minitransposons. When a parent full transposon has not been identified, only the inverted repeats of the proposed minitransposons are shown in red. The inverted repeats are indicated by red arrows (solid for the minitransposons, dotted for the full transposons). Flanking direct repeats, when present, are indicated by black arrows and are shown in black, underlined. SDR elements are highlighted according to the conventions of Figure 1. Underscores within the full transposons indicate open reading frames. The low order coordinates of each example, all from the chromosome of *N. punctiforme* except when otherwise indicated, are: (A1) 263683, (A2) 670380B, (A3) 3371581B, (A4) 2032965B, (B1) 697575B, (B2) 1598337B, (C1) 2898180B, (C2) 75121B in *Nodularia* CCY9414 Contig 003, (D1) 7174915B, (D2) 6179713F in the chromosome of *Anabaena variabilis*, (D3) 5233271F in the chromosome of *Anabaena* PCC 7120, and (D4) 114174B in plasmid pAvaC of *A. variabilis*. The letter after the coordinate indicates that the sequence shown lies on the forward (F) or backward (B) strand.

It is important to note that the SDR elements are generally distinct from the units in which they are contained. For example, SDR4 is more often found outside of the minitransposons described above, and the minitransposons are generally found without SDR elements. The heptameric repeats are found in the *Nostoc* genome far more frequently without SDR elements.

The flanking sequences for some SDR elements are highly variable (Supplemental Figs. 1–8). Except for transposon-borne SDR5.3, SDR5 elements generally do not share flanking sequences, and the same is true with SDR7. At the other extreme are SDR3 and SDR8 elements, almost all of which are flanked by the same tandem repeats (SDR3) or larger conserved sequence (SDR8). SDR4 lies in between these two extremes, as members of subfamilies within SDR4 share flanking sequences, but there is a great deal of variability in the flanking sequences of different subfamilies and often within subfamilies.

SDR1.17 (Fig. 1) is one of many subfamilies whose members have nearly identical flanking sequences that cannot be explained by appealing to heptameric repeats or transposable elements. These are discussed in the next section.

Nested SDR elements

All the SDR elements characterized in this study are smaller than 30 nt, but the sequence interpolated into the $tRNA^{leu}$ introns of *Nostoc* Nos41 and Nos33 are 48 and 45 nt in length, respectively

(Fig. 1, under SDR1 and SDR4). Both, however, can be viewed as one SDR element inserted into another: an unnamed SDR1 element inserted into an instance of SDR1.21 in the first case and SDR4.4 into SDR1.21 in the second case. Careful consideration of the flanking sequences of SDR elements reveals that such nested SDR elements are extremely common in *N. punctiforme*.

Figure 4A shows a typical nested SDR and how it might have arisen by successive insertions of SDR4 into itself, targeting the sequence CGAAG. Figure 4B shows a more extreme example, with six different insertions of SDR1 into a putative minitransposable element and provides a plausible sequence of events that could have given rise to the 11 apparently related minitransposon derivatives that are observed in the current genome of *N. punctiforme*. These sequences illustrate not only the ability of SDR1 to insert into new locations, but also the ability of degradative forces—deletion and mutation—to reclaim genome space taken up by repeated sequences. It seems fair to conclude that except in those cases where repeated sequences provide selective advantage, we see only those instances that have appeared relatively recently in evolutionary time.

Not all nested elements can be explained by a branching series of insertions and mutations. Figure 4C shows a set of related sequences that cannot be so explained, without resorting to unlikely coincidences, for example, insertion of the same SDR4/6 at the same position within two different instances of SDR1. The sequences point to a mechanism to spread a mutation at one site to other similar sequences, perhaps by recombination amongst chromosomal copies or by gene conversion.

Given the complexity of nested elements illustrated by Figure 4, it is clearly difficult to automate their identification; hence, we do not know the full extent to which nested elements occur in the genome. To get some idea, we performed an exhaustive search for SDR1 elements identified by the pattern ". . ..(<<<<)GAGCG.AG.CGA(>>>>)" (where <<<< >>>> indicates 4-nt inverted repeats) and split by insertions from 21 to 28 nucleotides. A total of 72 such cases were found in the *N. punctiforme* genome (45% of the number of unsplit instances of SDR1), of which 37% are interrupted by an insertion of SDR1 and 43% by SDR4. Looking at it from the inside out, of instances of SDR4 in subfamilies with three or more copies, 54% lie within either SDR1 or SDR4. In instances of SDR1, 27% lie within SDR1. Fig. 1 indicates some of the diversity of these nested elements (Supplemental Figs. 1 and 4 show all cases with high-copy subfamilies).

In contrast, SDR6 is part of what appears to be a single fused element with SDR4. A total of 87% of all instances of SDR6 are preceded by SDR4 (or its residue) in the same position and the same inverted orientation, and the same is true of instances in related cyanobacteria (Supplemental Fig. 6), suggesting a single origin. There is no clear evidence of an independent existence of SDR6. The SDR4/SDR6 unit lies within several different contexts, consistent with the idea that the fused unit is mobile. One insertion evidently occurred within SDR1.2 to form the initially anomalous SDR1.3, mentioned earlier.

The insertion of SDR4 into SDR1 indicates that SDR1 was present in the genome at a time when SDR4 was mobile. We looked for other such relationships and found insertions of SDR4 into SDR5 and the SDR4/SDR6 element as well as into itself (Supplemental Fig. 4), SDR1 into SDR5 as well as into itself (Supplemental Fig. 1), and the SDR4/SDR6 element into SDR1 as well as into itself (Supplemental Figs. 4, 6). We found no instances of other SDR elements participating in nesting. One might derive from this an order of appearance of SDR5, then SDR1, then SDR4/SDR6, and finally SDR4, but the true sequence of events is undoubtedly more complex, as will be discussed.

SDR elements in other cyanobacteria

In order to relate the presence of SDR elements to organismal phylogeny, we considered the genomes of the cyanobacteria shown in Figure 5. Except for SDR1 and SDR7, the SDR elements were confined to a coherent class, the heterocystous cyanobacteria, represented by *N. punctiforme*, *Anabaena* PCC 7120, *A. variabilis* ATCC 29,413, and *Nodularia* CCY9414. A few instances of SDR7 were also found in strains within a larger class, the filamentous cyanobacteria, of which the heterocystous cyanobacteria are a part, and one unicellular strain, *C. watsonii*, related to filamentous cyanobacteria, possesses many SDR1 elements. The original SDR elements in the tRNA intron from different strains of *Nostoc* in five of six cases match exactly the sequence of an SDR element from *N. punctiforme*, and in four of those instances, the flanking sequences also match.

It is of obvious interest whether apparent insertions of SDR elements are of ancient or recent origin. If the former, then the sites of insertion might be conserved amongst related organisms. We aligned and examined sequences at all informative sites for each element, an informative site defined as one occurring within a conserved gene (a gene with orthologs in both organisms under consideration) or between the same two conserved genes. The results of this analysis are summarized in Table 5 and discussed below.

SDR2a and SDR4 are the most common elements amongst the heterocystous cyanobacteria, with dozens of instances in all four of the sequenced genomes (Fig. 1; Supplemental Figs. 2A, 4), but they are quite different in the degree to which their insertion sites are shared amongst the genomes. Instances of SDR4 generally occur at sites that are unrelated from one organism to the next. Of 108 instances of SDR4 in *N. punctiforme* that are informative, none are shared in *Anabaena* PCC 7120, and the instances of SDR4 in each organism have different sets of characteristic flanking sequences. Even more telling, of 69 informative occurrences in *Anabaena* PCC 7120, only three are shared with the closely related strain *A. variabilis*. Most sites of SDR4 have evidently appeared or disappeared more recently than the divergence of the two species of *Anabaena*.

In contrast, instances of SDR2a often occur at the same sites from one organism to the next, indicating an origin more ancient than the divergence of the organism. Of 38 informative occurrences in *Anabaena* PCC 7120, 35 are shared with *A. variabilis*, and four of these are also shared with *N. punctiforme* (for example, see Fig. 6A). Those that are shared have exactly the same SDR2a sequence 48% of the time, far more frequently than can be accounted for by chance. *A. variabilis* has 150% more instances of SDR2a than *Anabaena* PCC 7120. It is not clear why this is the case. It is true that only *A. variabilis* has an apparently active transposase for the SDR2a-related minitransposon (Fig. 3D), but more than half of the SDR2a sites found in *A. variabilis* but not in *Anabaena* PCC 7120 are not obviously associated with a transposable element or derivative. Furthermore, there is no obvious difference between conserved and nonconserved SDR2a elements in *A. variabilis* with regard to either position (as described in Table 4) or flanking sequences.

SDR2a elements in both *Anabaenas* are often flanked by a heptameric repeat similar to STRR2 (Mazel et al. 1990; Meeks et

al. 2001) (GACAACT). Curiously, in the few instances where an instance of SDR2a is conserved between *Anabaena* and *Nostoc*, the *Anabaena* element is flanked by the STRR2 sequence, typical of *Anabaena* SDR2a elements, while the *Nostoc* element is flanked by the STRR8 sequence, typical of *Nostoc* SDR2a elements (see Fig. 6C). If, as is likely, the SDR2a elements at these sites share a common origin, then flanking sequences must somehow have become changed to accord with the tendencies of the genome as a whole.

Instances of SDR1 and SDR4/SDR6 follow the model of SDR4, not lying in conserved positions in the organisms in which they are found. SDR1 instances are seldom flanked by the same sequences found next to SDR1 elements from *Nostoc*. Eight SDR1 elements from *Nodularia* and eight from *Crocosphaera* are excep-

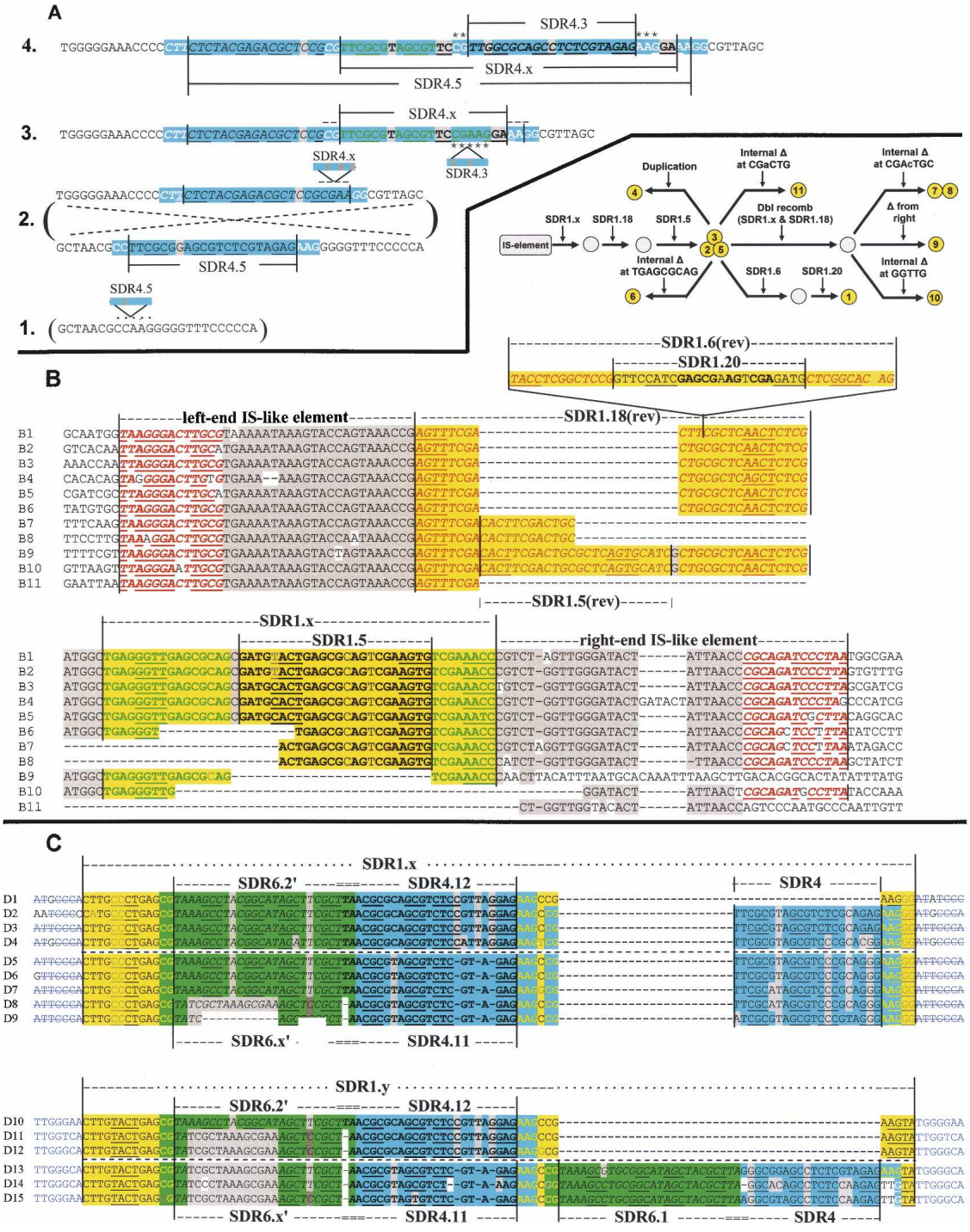
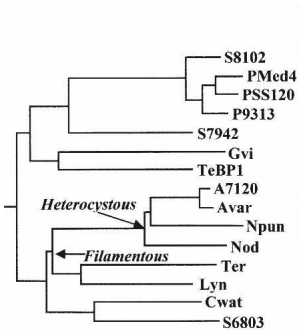


Figure 4. (Legend on next page)

tional in that their flanking sequences match the heptameric repeats found flanking the similar element SDR1.7 of *Nostoc*. Instances of nested SDR1 elements are found in both *Nodularia* and *A. variabilis*. This is also true for SDR4/SDR6.

SDR3 follows the model of SDR2a. All five informative instances of SDR3 in *Anabaena* PCC 7120 are shared with *A. variabilis*, while none are shared with *Nostoc*. Almost all of the sequences are flanked by at least a recognizable vestige of the same heptameric repeat that flanks SDR3 in *Nostoc*. Its insertions appear to predate the divergence of the two *Anabaenas*. Table 5 would seem to indicate an ancient origin of SDR5 insertions as well. Indeed, one insertion is found conserved in orthologs of NpF5954 amongst all four heterocystous cyanobacteria (Supplemental Fig. 12).

SDR2b is represented by only one instance in both strains of *Anabaena*, at a position conserved between the two, but not with *Nostoc*. None of the few instances of SDR7 and SDR8 in cyanobacteria beside *Nostoc* were in conserved positions.



Organism	Size (Mb)	Source	SDR Incidence 1	2a	2b	3	4	5	6	7	8
Synechococcus WH8102 [A]	2.43	JGI/Aug '03	~								
Prochlorococcus MED4 [B]	1.66	JGI/Aug '03									
Prochlorococcus SS120 [C]	1.75	GI:33239452									
Prochlorococcus MIT9313 [B]	2.41	JGI/Aug '03									
Synechococcus PCC 7942 [D]	2.75	JGI/Nov '04									
Gloeobacter violaceus [E]	4.66	Kazusa/Aug '03									
Thermosynechococcus BP1[F]	2.59	Kazusa/Aug '02									
Anabaena PCC 7120 [G]	7.21	Kazusa/Oct '01	~	+		~		~	+	+	~
Anabaena variabilis ATCC 29413	7.11	JGI/Jan '05	~	+		~		~	+	~	+
Nostoc punctiforme [H]	9.06	JGI/Feb '04	+	+		+		+	+	+	+
Nodularia CCY9414	5.32	JGI/Dec '06	+	+				+	~	+	~
Trichodesmium erythreum	7.76	GI:113473942									~
Lyngbya PCC 8106	7.04	GI:119483230									~
Crocosphaera watsonii	6.22	JGI/Feb '04	+								
Synechocystis PCC 6803 [I]	3.96	Kazusa/Dec '03									
Escherichia coli K12 [J]	4.64	GI:49175990	~								

Figure 5. Organisms used in cross-species comparisons of SDR elements. The phylogenetic tree to the *left* of the organism names is derived from a comparison of cyanobacterial 16S rRNA sequences. Arrows point to nodes defining the heterocystous cyanobacteria (the major focus of this work, highlighted in the table) and the more encompassing clade of filamentous cyanobacteria. All genomes except those from *Lyngbya* and *Crocosphaera* have been completely sequenced. The sequences were taken from Joint Genome Institute (http://genome.jgi-psf.org/mic_home.html), CyanoBase/Kazusa (http://bacteria.kazusa.or.jp/cyano/cyanobase/), or the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). In the latter case, the GenBank accession number is given. The small plasmid, pANS, of *Synechococcus* PCC 7942 was also taken from GenBank, accession GI:247785. The *Nostoc punctiforme* genome has recently been deposited in GenBank (accession no. NC_010628), with only minor differences in the annotation relative to the version used in this study. The incidence of each of SDR family (SDR1 through SDR8) is shown for each genome, represented by a plus (at least 10 occurrences), a tilde (at least one, but fewer than 10 occurrences), or nothing (0 occurrences). Published descriptions of the genomes: (*A*) Palenik et al. 2003, (*B*) Rocap et al. 2003, (*C*) Dufresne et al. 2003, (*D*) Holtman et al. 2005, (*E*) Nakamura et al. 2003, (*F*) Nakamura et al. 2002, (*G*) Kaneko et al. 2001, (*H*) Meeks et al. 2001, (*I*) Kaneko et al. 1996, (*J*) Blattner et al. 1997).

Unit lengths and target preferences of SDR elements

We compared conserved sites containing SDR elements in one genome but not in another, hoping thereby to determine their boundaries and to deduce what sequence features attract new insertions. Seven such insertions of SDR4 were found within conserved regions, as previously defined (Fig. 6F; Supplemental Fig. 12). In five of these instances, the insertion was coextensive with the proposed length of the SDR element. The observed alignments are more consistent with insertions rather than deletions, as in all three cases where an apparent insertion of SDR4 was found within a gene with orthologs in other cyanobacteria, alignments of the orthologs show gaps where SDR4 should appear. In all cases, the sequences flanking the insertion of SDR4 was consistent (at least three identities) with the 5-nt target sequence, CG|AAG, deduced from the analysis of nested elements (see above).

Of 12 informative and interpretable instances of SDR1 elements (Supplemental Fig. 12), most are explainable by recombination, using flanking heptameric repeats (e.g., Fig. 6A), but four appear to be insertions of SDR1 or a nested version of SDR1 (e.g., Fig. 6B). In all of those cases, the apparent insertion unit is 24 nt (or multiples of 24 in the case of nested elements). No target sequence is evident.

SDR5 shows an even stronger preference for a target sequence. Of 31 insertions of SDR5 into conserved regions of genes with orthologs in two organisms we examined, 29 showed insertions of 21 nt, the proposed length of SDR5 (Fig. 6G; Supplemental Fig. 12). Remarkably, the apparent target site of SDR5 is GCG|ATCGC, a sequence (called HIP1) that is highly overrepresented in many cyanobacteria (Robinson et al. 1995) and possibly involved in recombination (Robinson et al. 1997). As with SDR4, the events leading up to the presence or absence of SDR5 must be interpreted as insertions, not deletions, since orthologs of genes bearing SDR5 in nonheterocystous cyanobacteria lack the insert. Of the 67 instances of SDR5 elements in *N. punctiforme* with copy number three or greater, 82% lie within HIP1 sites between the third and fourth position, and all of the

Figure 4. Nested SDR elements. (*A*) A nested structure consisting of multiple SDR4 elements using color conventions described in Figure 1. The nested element (line 4), located in the chromosome of *N. punctiforme* from 671,496 to 671,583 (inverted), may have arisen, as shown, by the insertion of SDR4.5 (lines 1 and 2), followed by the insertion of an unnamed SDR4 element in the opposite orientation (lines 2 and 3), and finally by the insertion of SDR4.3 (lines 3 and 4). The sequence of the proposed intermediate shown in line 3 actually exists elsewhere in the chromosome, between coordinates 3,226,461 and 3,226,527 (inverted). Proposed intermediates 1 and 2 are hypothetical. (*B*) Nested structures derived from multiple apparent insertions of SDR1 elements, highlighted in yellow. SDR1 elements in the forward orientation are given in black or green characters. Those in the reverse orientation are given in red characters. Sequences in red characters flanking the nested elements are the inverted terminal repeats of a putative minitransposon into which the nested elements are embedded. Sequences highlighted in gray are conserved but otherwise uncharacterized. The low-order coordinates of the sequences shown are (B1) 3368609F, (B2) 2564863B, (B3) 1746187F, (B4) 8086669B, (B5) 633195F, (B6) 50757F, (B7) 8020729F, (B8) 6448446F, (B9) 1811789b, (B10) 7091736B, and (B11) 2217329F, where F and B indicate the forward and backward (inverse) strand, respectively. The *inset* at *top* presents a series of plausible events that might have given rise to the observed sequences. Numbers in circles refer to the lines in the alignment. Gray circles are hypothetical intermediates. (*C*) Seemingly related compound elements composed of SDR1, SDR4, and SDR6. The sequences are organized first by the outermost elements, two unnamed subfamilies of SDR1. Each grouping is then divided according to whether SDR4.11 or SDR4.12 has inserted into the SDR1 element. The low-order coordinates of the sequences shown are (D1) 3134527B, (D2) 2144806F, (D3) 4962007F, (D4) 3363464F, (D5) 5179224F, (D6) 8021963F, (D7) 6220474F, (D8) 733274B, (D9) 187174B, (D10) 676346B, (D11) 5070575B, (D12) 7201684F, (D13) 6371472B, (D14) 5596457F, and (D15) 3424319B.

Table 5. SDR elements in conserved sites

Family	Npun vs. A7120	Npun vs. Avar	Npun vs. Nod	A7120 vs. Avar
SDR1	(104) **0** (3)	(98) **0** (0)	(85) **0** (9)	(4) **0** (1)
SDR2a	(71) **4** (20)	(73) **5** (45)	(58) **4** (49)	(37) **34** (98)
SDR2b	(6) **0** (0)	(6) **0** (0)	(6) **0** (0)	(1) **1** (1)
SDR3	(19) **0** (3)	(17) **0** (2)	(15) **0** (9)	(5) **5** (5)
SDR4	(108) **0** (33)	(108) **0** (24)	(100) **1** (24)	(69) **3** (50)
SDR5	(37) **1** (7)	(39) **2** (4)	(33) **1** (9)	(5) **2** (6)
SDR6	(24) **0** (0)	(26) **0** (1)	(23) **0** (9)	(3) **0** (4)
SDR7	(8) **0** (0)	(8) **0** (0)	(16) **0** (0)	(0) **0** (0)
SDR8	(4) **0** (0)	(4) **0** (0)	(6) **0** (0)	(0) **0** (0)

The number in bold type is the number of instances the given SDR element is found in both indicated organisms at the same informative position, defined as either inside of a conserved gene or between the same two conserved genes. Genes are considered conserved if they are orthologs as described in the Methods. The first and second numbers in parentheses are the total numbers of informative instances of the given SDR element in the first organism and second organism, respectively.

remaining elements are flanked by a HIP1-like site (Fig. 1; Supplemental Fig. 5). Apart from a near absolute preference for A and T and positions 3 and 4, respectively, deviations from the canonical HIP1 sequence are equally likely at all positions (data not shown). SDR5 therefore has an extreme preference for HIP1 sites, one that is no less strong with instances in other heterocystous cyanobacteria. This is sufficient to explain the absence of SDR5 insertions into other SDR elements, as noted above, since HIP1 sites do not appear in other SDR elements.

SDR2a presents quite a different picture. There are 64 instances in which *A. variabilis* has a copy of SDR2a between two conserved genes, but *Anabaena* PCC 7120 does not in the corresponding position (Supplemental Fig. 12; data not shown). Alignments of these regions were carefully examined, and no instance was found showing clear evidence of a new insertion. Instead, 35% of the instances appeared mediated by transposition of the minitransposon that contains SDR2a.5 (Fig. 3D; Supplemental Fig. 2A). A total of 49% appeared to have arisen through recombination at sequences flanking the SDR element—most using variations on tandemly repeated AACAACT, including 10 events that seemed to be duplications of SDR2a a short distance from the first instance (Fig. 6D). Of the remaining instances, 6% had tandem instances of AACAACT flanking the SDRa element but no discernible heptamer in PCC 7120, and the rest could not be interpreted, generally because the intergenic region had diverged beyond recognition. In no case is there clear evidence of insertion except as mediated by known processes.

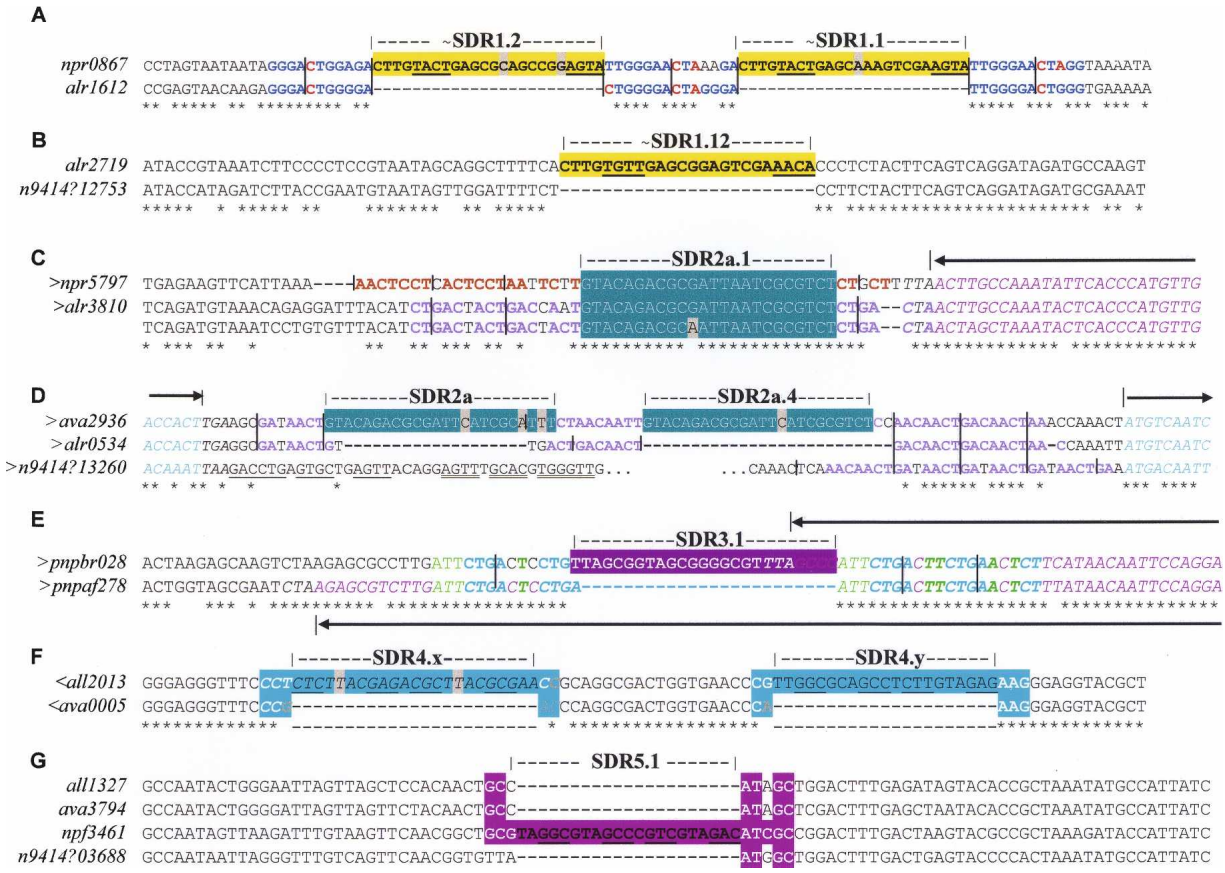


Figure 6. Sequence comparisons amongst cyanobacteria. Sequences of orthologous genes or regions adjacent to orthologous genes are shown where at least one of the sequences carries an SDR element. Color conventions are as described in Figure 1. In addition, gene sequences are shown italicized and in dark cyan (forward) or magenta (backward), and the extent of the reading frame is indicated by arrows. The sequences are taken from *N. punctiforme* (genes *npr0867, npf3461, npr5797, pnpbr028,* and *pnpaf278*), *Anabaena* PCC 7120 (genes *alr0534, all1327, alr1612, all2013, alr2719,* and *alr3810*), *Anabaena variabilis* (genes *ava0005, ava2936,* and *ava3794*), and *Nodularia* CCY9414 (genes *n9414?03688, n9414?12753,* and *n9414?13260*). Asterisks indicate positions of conservation amongst all sequences given.

There are only a few instances of SDR3 within *Nostoc* genes in regions sufficiently conserved to permit comparison with orthologous genes in *Anabaena* or vice versa (Fig. 6E; Supplemental Fig. 12). All interpretable cases where SDR3 was present in one genome but not the other could be explained by recombination between regions of tandem repeats. We found no instances of SDR2b, SDR6, SDR7, or SDR8 within regions sufficiently conserved to permit comparison of the site between genomes.

Discussion

Dispersed repeated sequences comprise a significant fraction of genomes in all branches of life, and they are fundamental to important evolutionary processes (Kazazian 2004; Lowe et al. 2007). The genomes of *N. punctiforme* and its relatives possess previously unreported repeated sequences in the range of from 21 to 27 nt, some of them evidently mobile. Since this is less than half the size of any sequence we know of with demonstrated mobility, a certain amount of skepticism is understandable. We have therefore reviewed below the evidence concerning each of the major claims of this work.

Are SDR elements mobile?

We consider an element to be mobile if it appears in new genome positions as a unit, if it does so independently of its immediate originating context, and if it relies on a mechanism distinct from recombination and gene conversion. The first condition excludes, for example, the 8-nt HIP1 sequences of cyanobacteria (Robinson et al. 1997) and the 9-nt transformation uptake signals of *Haemophilus influenzae* (Smith et al. 1999), which appear to achieve their high copy numbers by mutation in place, maintained by selection. The second condition excludes, for example, transposase genes, which are mobile only when flanked by the termini of their transposons. Also excluded are sequences that appear in new positions owing to recombination mediated by flanking repeated sequences (Kazazian 2004). The third condition prevents us from folding the flanking repeats of the last example into the units and calling them mobile elements.

By these criteria, there is strong evidence to support the claim that SDR1, SDR4, and SDR5 are mobile elements. All three (especially SDR5) are flanked by a diversity of flanking sequences (Fig. 1; Supplemental Figs. 1, 4, 5). Their insertion as a unit independent of recombination and gene conversion is evident from numerous comparisons of orthologous genes (Fig. 6; Supplemental Fig. 12). Furthermore, SDR1 and SDR4 often appear to have inserted into copies of themselves at various sites. The case for the fused SDR4/SDR6 element is weaker, relying on several instances where it appears to have inserted into other SDR elements (Fig. 4D; Supplemental Fig. 6).

In contrast, there is little evidence for the mobility of the other SDR elements. The many instances of SDR2a and fewer instances of SDR3 (Fig. 6; Supplemental Fig. 12) are either explained by recombination or are uninterpretable, and they have little diversity in flanking sequences (Fig. 1; Supplemental Figs. 2A, 3). SDR7 has a great deal of diversity in flanking sequences, but we have no evidence of insertion as a unit as opposed, for example, to arising by mutation in place.

Do SDR elements function as RNA?

This claim rests entirely on mutational analyses that in certain instances point to secondary structures apparently preserved by selection. The case is strong for SDR1, SDR4, and SDR5, where the number of compensatory mutations in regions of putative secondary structure far exceed expectation (Table 2). A weaker case can be made on this basis for SDR2a, one that is bolstered by the statistically significant excess of single transitions leading to potential G-T base pairs. The existence of base pairing in vivo within a single strand of double-stranded DNA is controversial (Kurahashi et al. 2004), and so we extend the inference of secondary structure to indicate a selectable role for RNA in SDR function. It is important to note that this role does not necessarily have to be in the propagation of the element. Small RNAs have been previously associated with certain repeated sequences in *Drosophila* (Aravin et al. 2003).

The propensity of SDR1 and SDR4 to insert themselves into other SDR sequences may be another piece of evidence pointing to an RNA intermediate. The insertions occur at multiple sites within the elements, indicating that it is not sequence that attracts the inserting element, but a more general property associated with target elements. The existence of SDR elements as RNA molecules, perhaps bound to protein, could set them apart from other sequences in the genome.

Can the mobility of SDR elements be explained by known mechanisms?

The mechanism or mechanisms by which SDR elements move are not clear. DNA-based transposition is not a likely explanation. Unlike most transposons (Siguier et al. 2006), SDR1, SDR4, SDR5, and SDR6 do not possess terminal inverted repeats and do not generate flanking direct repeats. If these SDR elements were minitransposons, they would be by far the smallest reported, and the lengths of the elements appear to be too small to fit within the topological constraints imposed by the known mechanisms of transposition (Lane et al. 1994; Gueguen et al. 2005).

Retrotransposition is a more appealing possibility. Genes capable of encoding proteins similar to reverse transcriptase within possibly mobile Type II introns have been observed in *N. punctiforme* (Dai et al. 2003; Doulatov et al. 2004) and other cyanobacteria (Nakamura et al. 2002; Dai et al. 2003; Doulatov et al. 2004). However, the apparently mobile SDR elements differ from LTR retrotransposons in that they lack long terminal repeats, and they differ from non-LTR retrotransposons in that they do not generate flanking direct repeats (Ostertag and Kazazian 2001). Type II introns (Lambowitz and Zimmerly 2004) are able to retrotranspose to ectopic locations in the genome, but the length of the apparent targets observed in these events are considerably greater than those of the mobile SDR elements (Dickson et al. 2001; Lambowitz and Zimmerly 2004; Fernández-López et al. 2005) and a degenerate type II intron the size of SDR5 would be less than one-fourth the size of any previously observed (Lambowitz and Zimmerly 2004).

A provocative clue as to mechanism may be provided by the nearly absolute target requirement of SDR5 for HIP1 sequences, GCGATCGC. The observed role of HIP1 in recombination (Robinson et al. 1997) is reminiscent of another 8-nt sequence, chi, that is part of the recombination process in *E. coli* (Smith 2001). Chi sites are recognized by RecBCD holoenzyme either as a proximal target for nicking or as a signal to stop DNA degradation. Either way, DNA near chi sites are rendered recombinogenic. If HIP1 sites serve a similar role, then base pairing between the exposed HIP1 DNA and the ends of SDR5 RNA may lead to inte-

gration of SDR5 through target-primed reverse transcription (Lambowitz and Zimmerly 2004).

SDR elements do not appear to be independently mobile. SDR5.1 in *N. punctiforme* occurs 30 times, each in a different context, but the identical sequence occurs only once in *Anabaena* PCC 7120. The difference may be that the *Nostoc* sequence occurs in a special transcriptional context, or *Nostoc* but not *Anabaena* may possess a hypothetical RNA-binding protein that facilitates the mobility of SDR5. The latter hypothesis is consonant with the common putative structures of SDR4, SDR5, and SDR6. Each possesses two potential GC-rich stems topped by loops of the pattern GyA (Fig. 2). Perhaps these serve as binding sites for a protein required for mobility of the elements (it is of course possible that they serve a functional role independent of mobility).

Are SDR elements similar to previously described repeated sequences?

The families of SDR elements in *Nostoc* exhibit different characteristics from previously reported small dispersed repeats. Unlike SDR elements, the previously described ERIC/IRU (De Gregario et al. 2005; Wilson and Sharp 2006), RSA (Bachellier et al. 1999), and BoxC (Bachellier et al. 1999) elements in enteric bacteria, RUP (Oggioni and Claverys 1999) and boxAB^{n}C (Martin et al. 1992; Knutson et al. 2006) elements in *Streptococcus*, and NEMIS (Mazzone et al. 2001) and SRE/Correia (Buisine et al. 2002) elements in *Neisseria* are all relatively long, imperfect palindromes possibly derived from transposons.

The PU/REP elements described in enterobacterial genomes (Bachellier et al. 1999) have characteristics with some similarity to SDR elements. They are short, 30–37 nt in length, with most of that length occupied by an imperfect palindrome. Like SDR1, SDR2a, and SDR2b, PU/REP elements end in an unpaired 4-nt tail, However, SDR2a and SDR2a augment their tails with nearly perfect palindromes while, at the other extreme, SDR1 has only four complementary nucleotides, surrounding a 12-nt conserved, nonpalindromic core. Like SDR1, but unlike SDR4, SDR5, and the SDR4/SDR6 fused element, PU/REP elements have no evident target site specificity. While they are frequently found in combination, as BIMEs (bacterial interspersed mosaic elements), they, unlike SDR1 and SDR4, have not been reported to form nested structures. There is no evidence we know of for the mobility of PU/REP elements.

What is the life cycle of an SDR element?

A global consideration of the findings may permit a partial reconstruction of the life history of SDR elements (Fig. 7). SDR elements must enter the lineage from within, by mutation, or from without, most likely by virus or incoming plasmid. In most cases, the new instance of the element suffers degradation and is lost from the organism, but occasionally it may become inserted within a tandomly repeated sequence or a transposon (full or miniature). Then it may be propagated even after it has ceased to be independently mobile (e.g., because the protein that acts on it is no longer present). This process is visible in instances of most of the SDR elements. Transposition continues until the gene encoding the transposase is inactivated by mutation.

Whether an SDR element arises by new insertion, transposition, or recombination (or gene conversion), it is bound to be lost by random mutation unless selection acts to preserve it. *Nostoc* and its relatives may have been subjected to waves of infection and recovery (Wagner 2006), by transposons and SDR elements, and we see only those infections in progress, scars of previous infections, and cases where random insertion has fortuitously placed an element in a context where it provides the organism with a selectable advantage. The conserved instances of SDR2a in *Anabaena* may be such cases, and the long inverted repeat provided by many of these instances downstream from genes may serve in transcriptional termination (as shown also by Lupski and Weinstock 1992 in some instances of PU/REP) or translational coupling. The less-common conserved upstream instances of SDR2a may stabilize transcripts, as has been shown for a boxAB^{n}C sequence in *Streptococcus pneumoniae* (Knutson et al. 2006).

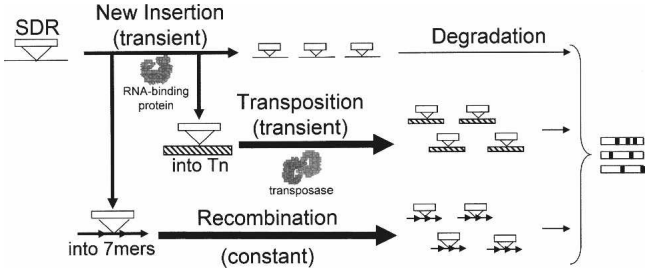


Figure 7. Model for propagation of SDR elements. The model follows the life cycle of an SDR element (white box). When a putative RNA-binding protein is present within the organism, the element is mobile and may insert into tandemly repeated sequences (tandem arrows), transposable elements (hatched box), or uncharacterized sequences (thin line). In the first two cases, the element may propagate to other sites in the genome even after the SDR element is no longer mobile. In the absence of selection, mutations (black boxes) accumulate in the element over time, giving rise to the diverse sequences observed in genomic sequences. See text for further explanation.

Why invent a new name for families of dispersed repeats?

There have been many names given in the past to small dispersed repetitive elements. Many have incorporated into the name the genus of the organism, for example, RUP (repeat unit of *Pneumococcus*) (Oggioni and Claverys 1999) and NEMIS (*Neisseria* miniature insertion sequence) (Mazzone et al. 2001). Others have used the name to describe the structure of the elements. The PU (palindromic unit) elements of *E. coli* and its relatives is a case in point. Neither convention seemed suitable for the elements described in this report, as they may be found in different genera and do not share a common structure. SRE (small repetitive elements) (Buisine et al. 2002) is less restrictive, but we would like to maintain a distinction between tandem repeats and repeats well separated on the genome. We have therefore opted for a name that captures only the essential characteristics—small dispersed repeats (SDR)—and have used it as a prefix for specific families, as Tn is used for transposon names.

Methods

Sequences

The tRNA^{Leu} (UAA) intron sequences used in the present study were recorded from strains of *Nostoc* associated with the lichens *Nephroma resupinatum* (L.) Ach., AF055660 (Nos30); *Peltigera britannica* (Gylen.) Holtan-Hartwig & Tønsb, AF176600 (Nos37); *P. venosa* (L.) Hoffm., AF176604 (Nos38); *Pimelea venosa*, AF176596 (Nos41), and the bryophytes *Anthoceros fusiformis* Aus-

tin, AF151776 (Nos51) and *A. fusiformis*, AF151779 (Nos54). The genomic sequences used in this study are referenced in Figure 5.

Database search

Searches and analyses were done within CyanoBIKE (http://biobike.csbc.vcu.edu), an instance of BioBIKE (Biological Integrated Knowledge Environment, formerly BioLingua) (Massar et al. 2004), an integrated knowledge-base and programming environment that facilitates genomic analysis. BLAST (Altschul et al. 1997) was used initially but was abandoned because its algorithm, requiring contiguous-word matches, failed to find many matches we deemed significant. Families of repeated sequences were found instead by iterative searches using a BioBIKE function to find all sequences related to an initial query (taken from the intron sequences) with fewer than three mismatches (or two mismatches in the case of the shorter SDR4 and SDR5 sequences), and then all sequences related to those found in the initial search. This level of similarity was chosen because by using it there is no more than a calculated 0.01% chance of finding a match in a random sequence the size and nucleotide composition of the *Nostoc* genome. Other genomes were searched in the same way, but using as initial queries all versions of SDR elements with at least three copies in *Nostoc*. In addition, SDR1 was sought in genomes by the less-stringent pattern match, as described in the text. One would expect a pattern match to arise in a random genome the size and composition of *E. coli's* about one time in four. Sequences matching a given pattern were found using BioBIKE's pattern-matching capabilities.

Sequences were aligned using ClustalW (Thompson et al. 1994) implemented within BioBIKE and also by hand. Highlighting of sequences to display their internal structure was done by hand. Information content at each position of a collection of SDRs was represented by WebLogos (Crooks et al. 2004). The SDRs were first visually filtered to remove any that were interrupted by the insertion of other SDR elements.

An exhaustive catalog of 24-nt repeated sequences was obtained by sorting all 24-nt sequences extracted from the genome, counting the occurrences of identical sequences, and sorting the counts.

Files containing sequences and coordinates of SDR elements are available from the website: http://www.people.vcu.edu/~elhaij/SDR.

Other methods

SDR elements from different cyanobacteria were compared by aligning sequences of conserved regions. Conserved regions were determined using orthologous genes as points of reference. We defined two genes from different organisms to be orthologous if each gene was most similar to the other in BLASTs of one gene against the genes of the other's organism, using a threshold of 10^{-10}. Finding orthologs were greatly facilitated by precomputed BLAST results within BioBIKE for all available cyanobacterial genomes.

The phylogenetic tree shown in Figure 5 was obtained from a distance matrix using PHYLIP, version 3.2 (http://evolution.genetics.washington.edu/phylip.html) as implemented in BioBIKE.

Acknowledgments

This study was financially supported by the National Science Foundation through its Bioinformatics and Bioengineering Summer Institute (M.K.), The Royal Swedish Academy of Sciences (J.-L.C.), and the Swedish Research Council (P.L.). We thank Frank Larimer (Oak Ridge National Laboratory) for access to the BLAST facility for *N. punctiforme* prior to public release, J.P. Massar and Jeff Shrager (BioBIKE support staff) for help in implementing search algorithms (now part of the language), and Franz, Inc. for allowing the use of Allegro Common Lisp that underlies BioBIKE.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Arakawa, K., Uno, R., Nakayama, Y., and Tomita, M. 2007. Validating the significance of genomic properties of Chi sites from the distribution of all octamers in *Escherichia coli*. *Gene* **392:** 239–246.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschi, T. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5:** 337–350.

Bachellier, S., Clément, J.-M., and Hofnung, M. 1999. Short palindromic repetitive DNA elements in enterobacteria: A survey. *Res. Microbiol.* **150:** 627–639.

Blattner, F.R., Plunket III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Buisine, N., Tang, C.M., and Chalmers, C. 2002. Transposon-like Correia elements: Structure, distribution and genetic exchange between pathogenic *Neisseria* sp. *FEBS Lett.* **522:** 52–58.

Costa, J.L., Paulsrud, P., and Lindblad, P. 2002. The cyanobacterial $tRNA^{Leu}$ (UAA) intron: Evolutionary patterns in a genetic marker. *Mol. Biol. Evol.* **19:** 850–857.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. 2004. WebLogo: A sequence logo generator. *Genome Res.* **14:** 1188–1190.

Dai, L., Toor, N., Olson, R., Keeping, A., and Zimmerly, S. 2003. Database for mobile group II introns. *Nucleic Acids Res.* **31:** 424–426.

De Gregario, E., Silvestro, G., Petrillo, M., Carlomagno, M.S., and Di Nocera, P.P. 2005. Enterobacterial repetitive intergenic consensus sequence repeats in *Yersiniae*: Genomic organization and functional properties. *J. Bacteriol.* **187:** 7945–7954.

de Hoon, M.J.L., Makita, Y., Nakai, K., and Miyano, S. 2005. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comp. Biol. Comput.* **1:** e5. doi: 10.1371/journal.pcbi.0010025.

Dickson, L., Huang, H.-R., Liu, L., Matsuura, M., Lambowitz, A., and Perlman, P.S. 2001. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proc. Natl. Acad. Sci.* **98:** 13207–13212.

Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., Simons, R.W., Zimmerly, S., and Miller, J.F. 2004. Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431:** 476–481.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M.Y., Koonin, E.V., Le Gall, F., et al. 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl. Acad. Sci.* **100:** 9647–9649.

Fernández-López, M., Muñoz-Adelantado, E., Gillis, M., Willems, A., and Toro, N. 2005. Dispersal and evolution of the *Sinorhizobium meliloti* group II RmInt1 intron in bacteria that interact with plants. *Mol. Biol. Evol.* **22:** 1518–1528.

Godde, J.S. and Bickerton, A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* **62:** 718–729.

Gueguen, E., Rousseau, P., Duval-Valentin, G., and Chandler, M. 2005. The transpososome: Control of transposition at the level of catalysis. *Trends Microbiol.* **13:** 543–549.

Holtman, C.K., Chen, Y., Sandoval, P., Gonzales, A., Nalty, M.S., Thomas, T.L., Youderian, P., and Golden, S.S. 2005. High-throughput functional analysis of the *Synechococcus elongatus* PCC 7942 genome. *DNA Res.* **12:** 103–115.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109–136.

Kaneko, T., Nakamura, Y., Wolk, C.P., Kuritz, T., Sasamoto, S., Watanabe,

A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., et al. 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res.* **8:** 205–213.
Kazazian, H.H. 2004. Mobile elements: Drivers of genome evolution. *Science* **303:** 1626–1632.
Knutson, E., Johnsborg, O., Quentin, Y., Claverys, J.-P., and Håvarstein, L.S. 2006. BOX elements modulate gene expression in *Streptococcus pneumoniae*: Impact on fine-tuning of competence development. *J. Bacteriol.* **188:** 8307–8312.
Kurahashi, H., Inagaki, H., Yamada, K., Ohye, T., Taniguchi, M., Emanuel, B., and Toda, T. 2004. Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. *J. Biol. Chem.* **279:** 35377–35383.
Lambowitz, A.M. and Zimmerly, S. 2004. Mobile group II introns. *Annu. Rev. Genet.* **38:** 1–35.
Lane, D., Cavaillé, J., and Chandler, M. 1994. Induction of the SOS response by IS*1* transposase. *J. Mol. Biol.* **242:** 339–350.
Lowe, C.B., Bejerano, G., and Haussler, D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci.* **104:** 8005–8010.
Lupski, J.R. and Weinstock, G.M. 1992. Short, interspersed repetitive DNA sequences in prokaryotic genomes. *J. Bacteriol.* **174:** 4525–4529.
Martin, B., Humbert, O., Camara, M., Guenzi, E., Walker, J., Mitchell, T., Andrew, P., Prudhomme, M., Alloing, G., Hakenbeck, R., et al. 1992. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. *Nucleic Acids Res.* **20:** 3479–3483.
Massar, J., Travers, M., Elhai, J., and Shrager, J. 2004. BioLingua: A programmable knowledge environment for biologists. *Bioinformatics* **21:** 199–207.
Mazel, D., Houmard, J., Castets, A.M., and Tandeau de Marsac, N. 1990. Highly repetitive DNA sequences in cyanobacterial genomes. *J. Bacteriol.* **172:** 2755–2761.
Mazzone, M., De Gregorio, E., Lavitola, A., Pagliarulo, C., Alifano, P., and Di Nocera, P.P. 2001. Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic *Neisseriae*. *Gene* **278:** 211–222.
Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., and Atlas, R. 2001. An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynth. Res.* **70:** 85–106.
Mira, A., Ochman, H., and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17:** 589–596.
Mrázek, J., Gaynon, L.H., and Karlin, S. 2002. Frequent oligonucleotide motifs in genomes of three streptococci. *Nucleic Acids Res.* **30:** 4216–4221.
Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., et al. 2002. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res.* **9:** 123–130.
Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K., Kishida, Y., et al. 2003. Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res.* **10:** 137–145.
Oggioni, M.R. and Claverys, J.P. 1999. Repeated extragenic sequences in prokaryotic genomes: A proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology* **145:** 2647–2653.
Ostertag, E. and Kazazian, H.H. 2001. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35:** 501–538.
Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E.E., McCarren, J., et al. 2003. The genome of a motile marine *Synechococcus*. *Nature* **424:** 1001–1002.
Petrillo, M., Silvestro, G., Di Nocera, P.P., Boccia, A., and Paolella, G. 2006. Stem-loop structures in prokaryotic genomes. *BMC Genomics* **7:** 170. doi: 10.1186/1471-2164-7-170.
Rivas, E. and Eddy, S.R. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2:** 8. 10.1186/1471-2105-2-8.
Robinson, N.J., Robinson, P.J., Gupta, A., Bleasby, A.J., Whitton, B.A., and Morby, A.P. 1995. Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Res.* **23:** 729–735.
Robinson, P.J., Cranenburgh, R.M., Head, I.M., and Robinson, N.J. 1997. HIP1 propagates in cyanobacterial DNA via nucleotide substitutions but promotes excision at similar frequencies in *Escherichia coli* and *Synechococcus* PCC 7942. *Mol. Microbiol.* **24:** 181–189.
Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R., et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042–1047.
Rousset, F., Pélandakis, M., and Solignac, M. 1991. Evolution of compensatory substitutions through G-U intermediate state in *Drosophila* rRNA. *Proc. Natl. Acad. Sci.* **88:** 10032–10036.
Schneider, T.D. 2000. Evolution of biological information. *Nucleic Acids Res.* **28:** 2794–2799.
Shapiro, J.A. 2005. A 21st century view of evolution: Genome system architecture, repetitive DNA, and natural genetic engineering. *Gene* **345:** 91–100.
Siguier, P., Filée, J., and Chandler, M. 2006. Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* **9:** 526–531.
Smith, G.R. 2001. Homologous recombination near and far from DNA breaks: Alternative roles and contrasting views. *Annu. Rev. Genet.* **35:** 243–274.
Smith, H., Gwinn, M., and Salzberg, S. 1999. DNA uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.* **150:** 603–616.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Tobes, R. and Ramos, J.-L. 2005. REP code: Defining bacterial identity in extragenic space. *Environ. Microbiol.* **7:** 225–228.
Tóth, G., Gáspári, Z., and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10:** 967–981.
van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. 1998. Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.* **62:** 275–293.
Wagner, A. 2006. Periodic extinctions of transposable elements in bacterial lineages: Evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.* **23:** 723–733.
Wilson, L.A. and Sharp, P.M. 2006. Enterobacterial repetitive intergenic consensus (ERIC) sequences in *Escherichia coli*: Evolution and implications for ERIC-PCR. *Mol. Biol. Evol.* **23:** 1156–1168.
Yarnell, W.E. and Roberts, J.W. 1999. Mechanism of intrinsic transcription termination and antitermination. *Science* **284:** 611–615.
Zhou, F., Tran, T., and Xu, Y. 2008. *Nezha*, a novel active miniature inverted-repeat transposable element in cyanobacteria. *Biochem. Biophys. Res. Commun.* **365:** 790–794.

Received December 3, 2007; accepted in revised form May 12, 2008.