ISSN 0378-1119
Volume 473, Issue 2, 1 March 2011
ELSEVIER
GENE
An International Journal of Functional and Evolutionary Genomics

Gene 473 (2011) 139–149

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators

Shen Lin [a,b,c], Stefan Haas [b], Tomasz Zemojtel [b], Peng Xiao [a,b,c], Martin Vingron [b], Renhui Li [a,*]

[a] Key Laboratory of Aquatic Biodiversity and Conservation Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China
[b] Department of Computational Molecular Biology, Max Planck Institut für Molekulare Genetik, 14195 Berlin, Germany
[c] Graduate University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Accepted 26 November 2010
Available online 13 December 2010

Received by I. King Jordan

Keywords:
Transposable element
Insert sequence
IS diversity
Cyanobacterial genomes
IS family
IS subfamily

ABSTRACT

Transposable elements are widely distributed in archaea, bacteria and eukarya domains. Considerable discrepancies of transposable elements in eukaryotes have been reported, however, the studies focusing on the diversity of transposable element systems in prokaryotes were scarce. Understanding the transposable element system in cyanobacteria by the genome-wide analysis will greatly improve the knowledge of cyanobacterial diversity. In this study, the transposable elements of seventeen cyanobacterial genomes were analyzed. The abundance of insertion sequence (IS) elements differs significantly among the cyanobacterial genomes examined. In particular, water bloom forming *Microcystis aeruginosa* NIES843 was shown to have the highest abundance of IS elements reaching 10.85% of the genome. IS family is a widely acceptable IS classification unit, and IS subfamily, based on probe sequences, was firstly proposed as the basic classification unit for IS element system therefore both IS family and IS subfamily were suggested as the two hierarchical units for evaluating the IS element system diversity. In total, 1980 predicted IS elements, within 21 IS families and 132 subfamilies, were identified in the examined cyanobacterial genomes. Families IS4, IS5, IS630 and IS200-605 are widely distributed, and therefore supposed to be the ancestral IS families. Analysis on the intactness of IS elements showed that the percentage of the intact IS differs largely among these cyanobacterial strains. Higher percentage of the intact IS detected in the two hot spring cyanobacterial strains implied that the intactness of IS elements may be related to the genomic stabilization of cyanobacteria inhabiting in the extreme environments. The frequencies between IS elements and miniature inverted-repeat transposable elements (MITEs) were shown to have a linear positive correlation. The transposable element system in cyanobacterial genomes is of hypervariabilty. With characterization of easy definition and stability, IS subfamily is considered as a reliable lower classification unit in IS element system. The abundance of intact IS, the composition of IS families and subfamilies, the sequence diversity of IS element nucleotide and transposase amino acid are informative and suitable as the indicators for studies on cyanobacterial diversity. Practically, the transposable system may provide us a new perspective to realize the diversity and evolution of populations of water bloom forming cyanobacterial species.

1. Introduction

Transposable elements (also called mobile element or jumping genes) are widely distributed in a variety of organisms including prokaryotes and eukaryotes (Lepetit et al., 2002). A large amount of transposable elements enhanced the potential for their hosts' adaptation to different environments and created considerable interspersed repeats within genomes by transposition events accumulating over evolutionary time [(Nekrutenko and Li, 2001) and (Kidwell and Lisch, 2001)]. Transposable element system has been proven to be a powerful marker for divergent populations in different groups of organisms [(Lepetit et al., 2002), (Zampicinini et al., 2004), (Barnes et al., 2005) and (Boulesteix et al., 2007)]. In eukaryotic organisms, much is known about the transposable element system, including the element structure, transposition mechanisms, copy number variance (CNV) and evolutionary history of transposable elements [(Wicker et al., 2007) and (Langdon et al., 2003)]. In bacteria, insert sequences (IS) and miniature inverted-repeat transposable elements (MITEs) are two principal types of transposable elements, which can move from place to place via a DNA intermediate by a cut and paste mechanism (class II element) (Gray, 2000) or spread to other organisms by horizontal gene transfer [(Kidwell, 1992) and (Leavis et al., 2007)]. Insertion sequences in

Abbreviations: IS, insertion sequence; MITEs, miniature inverted-repeat transposable elements; CNV, copy number variance; ORFs, open reading frames; K2P, Kimura's two-parameter; NJ, neighbor-joining.

* Corresponding author. Tel.: +86 27 68780067; fax: +86 27 68780123.
E-mail addresses: biolin@ihb.ac.cn (S. Lin), haas@molgen.mpg.de (S. Haas), zemojtel@molgen.mpg.de (T. Zemojtel), xp@ihb.ac.cn (P. Xiao), vingron@molgen.mpg.de (M. Vingron), reli@ihb.ac.cn (R. Li).

0378-1119/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2010.11.011

prokaryotes were assumed to be an important driving force for novel genotypic and phenotypic variants. An investigation on the IS diversity of *Enterococcus faecium* confirmed that divergent IS could be used to distinguish subspecies from different environments and evaluated their evolutionary relationship (Leavis et al., 2007). Studies on the *Rhizobium meliloti* populations indicated IS-fingerprinting approach was a fine resolution for differing close species (strains) and would be suitable for ecological studies of individual strains in some complex ecosystem [(Kosier et al., 1993) and (Niemann et al., 1997)]. In addition, the evolutionary dynamics of insertion sequences in *Rhizobium etli* populations were shown to be related to the evolutionary histories of the chromosome and symbiotic plasmid (Lozano et al., 2010).

The recent release of prokaryotic genomes considerably contributed to the reorganization of a large number of IS families, especially in archaea. A systematical IS element collection and IS family based classification system have been established by some professional databases, such as IS Finder (Siguier et al., 2006) and GenBank. Cyanobacteria, considered as the ancestor of photosynthetic organisms on the earth, consist of large groups of organisms from unicellular to filamentous forms (Mulkidjanian et al., 2006). However, less is known about the transposable elements in cyanobacteria. IS elements have been briefly described in several cyanobacterial genomes [(Kaneko et al., 1996, 2001, 2007), (Nakamura et al., 2002) and (Nakamura et al., 2003)], and MITE was firstly analyzed in the recently released *Microcystis aeruginosa* NIES 843 genome. Zhou et al. (2008) reported the genetic map of recently active IS elements in cyanobacterial genomes, and they presented a heavy dependence of the activities of IS elements on the environments, and the close linkage between the abundance of recently active IS elements with genome size. However, recently released cyanobacterial genomes were not included in the above study, especially lacking high IS containing cyanobacterial genomes, which did not demonstrate and provide the general knowledge of IS diversity in cyanobacteria. Building a refine hierarchy for IS classification system is one goal of this study. IS family has been widely used in previous studies [(Kaneko et al., 2007), (Filée et al., 2007) and (Brügger et al., 2002)] and therefore recognized as an approved classification unit. However, the lower unit below IS family is obscure. IS group, a lower unit, was proposed and partly applied in the IS Finder database and in the comparative analyses on archaeal genomes by Chandler et al. [(Siguier et al., 2006) and (Filée et al., 2007)], but it is not easy to practically apply this IS group system because of its vague classification criterion, and incomplete database group annotation. Due to an extremely high diversity of IS nucleotide/transposase existing in prokaryotes, establishing a lower IS classification unit is highly expected. Therefore, IS subfamily, a new classification unit was suggested in this study.

In the present study, we analyzed and compared the general characters of transposable element systems in seventeen cyanobacterial genomes, including their abundance, distribution and family/subfamily compositions. Analyses on parsimonious evolutionary scenario, IS copy number variance, element intactness and the nucleotide and transposase amino acid sequences of these cyanobacterial transposable element systems, were performed as well. The framework for selecting the interspersed repeats encoding transposase was developed, and several complete cyanobacterial genomes released recently, including those from water bloom forming species such as *M. aeruginosa* NIES 843, *M. aeruginosa* PCC7806, *Trichodemium erythraeum* ISM101, as well the recently released cyanobacterial genome of *Cylindrospermopsis raciborskii* CS-505, which is also a frequently reported toxic bloom forming cyanobacterium in these years for its producing cylindrospermopsin [(Wilson et al., 2000) and (Stucken et al., 2010)], were included in this study. This combination is expected to achieve a comprehensive evaluation on the genetic diversity of cyanobacterial transposable system in more details and shed light on the feasibility of using the transposable element diversity information for the studies on cyanobacterial population diversity and evolutionary history.

2. Materials and methods

2.1. Genomes of cyanobacterial strains

Seventeen cyanobacterial chromosome genomes and plasmid sequences were used in this study, and these strains cover twelve genera with chromosome size from 1.68 Mbp to 8.23 Mbp. Besides the well sequenced and spliced ring shape genomes, some genomes are assemblages of contigs. The contig numbers of the genomes of *M. aeruginosa* PCC7806*, Raphidiopsis brookii* D9 and *C. raciborskii* CS-505 are 116, 47 and 93 respectively. The cyanobacterial strains used in this study can be morphologically divided into unicellular and filamentous, and have diverse inhabits including terrestrial, freshwater, marine water and hot spring (Table 1).

2.2. Construction of the nucleotide and transposase amino acid probe libraries

Two sets of IS sequence probe libraries (also called template library in some other studies) were generated in this research. The nucleotide probe library aims at rough nucleotide sequence mining, and the other was the transposase amino acid probe library corresponding to each nucleotide probe aiming at nucleotide candidate sequences reexamination and intactness judgment. The procedure for nucleotide probe library construction was as follows: all the repeat elements longer than 500 bp were collected using the Vmatch program package (Kurtz, 1999). Sequence consensus was executed by Cap3 program (Huang and Madan, 1999), and all the consensus sequences were examined by reiterative BLAST analysis setting the parameters of e value cutoff of 10^{-20} and key word of 'Transposase'. The positive hits of nucleotide sequences were selected as IS nucleotide probes. For transposase amino acid probes, the open reading frames (ORFs) of transposable element corresponding to each IS nucleotide probes were recognized by getorf program from the EMBOSS package. The longer ORF sequences as the best representative of the intact transposase corresponding to each nucleotide probe, were collected as IS transposase amino acid probes. The strategy used to define the ORFs in this study is searching the region that is free of STOP codons. IS family was identified by the homologous search mainly according to IS Finder and GenBank.

2.3. IS element mining

To identify possible IS elements in cyanobacterial genomes, each of genome sequences was screened with RepeatMasker 3.2.9 (Smit et al., 1996–2004), which is able to identify copies of IS element candidates by pairwise sequence comparisons with a self-constructive IS nucleotide probe library described above. The following arguments were used for this search: 'cross_match' as the search engine; 'slow' to obtain a search 0–5% more sensitive than default; 'nolow' to not mask low complexity DNA or simple repeats. All the nucleotide sequences screened out were regarded as IS candidates. The putative ORFs of these IS candidates recognized by EMBOSS: getorf were compared with amino acid probe library of the IS transposase by Blastp and the hits with lower e values ($1e^{-50}$) were picked out and recognized as the predicted IS elements. All the nucleotide sequences fished by the same nucleotide probe were classified into one subfamily. The reliability of this method is verified to be credible (Supplemental File 1).

Corresponding to the two sets of probe libraries above, two types of intact IS elements were defined (Fig. 1). N-intact elements represent ISs which cover at least 95% nucleotide sequence corresponding to the nucleotide probe. The ISs, which cover at least 99% amino acid sequence with correspondence to transposase amino acid probe, are defined as P-intact elements.

Table 1
Cyanobacterial strains used in this study and their genome information.

Species	GenBank No.	Habitat	Morphology	Length (nt)	GC%	Topology	Sequencing center	Released date
Microcystis aeruginosa NIES-843	AP009552	Freshwater lake	unicellular	5,842,795	42	circular	Kazusa, Japan	2008-1-31
Microcystis aeruginosa PCC 7806	AM778843-AM778958	Freshwater lake	unicellular	5,172,804	42	contigs	Institut Pasteur, France	2007-11-1
Synechocystis sp. PCC 6803	BA000022	Freshwater lake	unicellular	3,573,470	47	circular	Kazusa, Japan	2001-10-23
Synechococcus sp. JA-3-3Ab	CP000239	Hot spring	unicellular	2,932,766	60	circular	CAG, US	2006-2-7
Synechococcus elongatus PCC 7002	CP000951	Freshwater lake	unicellular	3,008,047	49	circular	Beijing Genomic Institute, China	2008-3-17
Trichodesmium erythraeum IMS101	CP000393	Marine	filamentous, non-heterocystous	7,750,108	34	circular	DOE	2006-8-30
Nostoc punctiforme PCC 73102	CP001037	Terrestrial	filamentous, heterocystous	8,234,322	41	circular	DOE	2008-4-25
Anabaena variabilis ATCC 29413	CP000117	Terrestrial	filamentous, heterocystous	6,365,727	41	circular	DOE	2005-9-20
Anabaena sp. PCC 7120	BA000019	Terrestrial	filamentous, heterocystous	6,413,771	41	circular	Kazusa, Japan	2001-11-28
Acaryochloris marina MBIC11017	CP000828	Marine	unicellular	6,503,724	47	circular	TGen Sequencing Center, US	2007-10-17
Cyanothece sp. PCC 7425	CP001344	Marine	unicellular	5,374,574	50	circular	DOE	2009-1-15
Prochlorococcus marinus str. MIT 9211	CP000878	Marine	unicellular	1,688,963	38	circular	MOORE	2007-11-13
Prochlorococcus marinus str. MIT 9215	CP000825	Marine	unicellular	1,738,790	31	circular	DOE	2007-9-21
Thermosynechococcus elongatus BP-1	BA000039	Hot spring	unicellular	2,593,857	53	circular	Kazusa, Japan	2002-8-19
Gloeobacter violaceus PCC 7421	BA000045	Terrestrial	unicellular	4,659,019	61	circular	Kazusa, Japan	2003-10-6
Cylindrospermopsis raciborskii CS-505	ACYA00000000	Freshwater lake	filamentous, heterocystous	3,879,030	40	contigs	Germany	2010-1-4
Raphidiopsis brookii D9	ACYB00000000	Freshwater lake	filamentous, non-heterocystous	3,186,511	40	contigs	Germany	2010-1-4

DOE means DOE Joint Genome Institute, US; MOORE means The Gordon and Betty Moore Foundation Marine Microbiology Initiative, US; NARA means Nara Institute of Science and Technology, Japan; CAG means Center for the Advancement of Genomics, US.

2.4. MITE element mining

The strategy for the MITE search is an integration of repeated elements and TIR/DR border identification. All the repeated elements longer than 100 bp were collected by the Vmatch package, and 15 bp left/right flanking wings were added to ensure the potential intactness of TIR/DR border. The candidates containing the TIR/DR structure and shorter than 499 bp by MUST (Chen et al., 2009) were defined as MITE. The genomes were scanned using RepeatMasker with the same argument setting to IS mining, and all the sequences homologous to the nucleotide probes were defined as type I, and the remains were type II.

2.5. Phylogenetic analysis

Nucleotide and amino acid sequences were aligned using either CLUSTALW, version 2.0 (Larkin et al., 2007) or MUSCLE (Edgar, 2004). Genetic distances were calculated using the method of Kimura's two-parameter (K2P) for DNA sequences and Poisson correction for protein sequences. The phylogenetic trees were constructed from the multiple-aligned data using the neighbor-joining (NJ) algorithmic. Kimura's two-parameter was implemented within the MEGA4 program package (Tamura et al., 2007).

3. Results

3.1. Abundance and basic properties of cyanobacterial IS

Totally 1980 predicted IS elements including intact and fragmentary ones, were detected in these cyanobacterial genomes (Supplemental File 2), and the abundance of the predicted ISs in different strains varies considerably. *M. aeruginosa* NIES 843, a unicellular water bloom forming strain with the genome size as 5.8 Mbp, showed to contain the highest IS abundance in the examined strains as 532 IS elements, covering 10.85% of the genome (Figs. 1, 2 and Table 2). While another *M. aeruginosa* PCC 7806 strain was revealed to have 359 pieces of IS elements with 8.98% coverage of the genome. Strains *Acaryochloris marina* MBIC11017 and *Thermosynechococcus elongatus* BP-1 were presented to have the IS coverage over 3%. Surprisingly, none of the IS elements were detected in two marine strains *Prochlorococcus* sp. MIT 9211 and *Prochlorococcus* sp MIT 9215. The length of the predicted IS elements ranged from 199 bp to 6495 bp, with the majority within the range of 500–2750 bp (Supplemental Fig. S2). A small amount of IS elements longer than 3 kb were also detected, including the elements from *M. aeruginosa* PCC7806, and Tn elements longer than 4 kb from *A. marina* MBIC11017, *Nostoc punctiforme* PCC 73102 and *Anabaena variabilis* ATCC 29413. One IS element could be detected as roughly 45 kb size within the cyanobacterial genome. *Trichodesmium erythraeum* IMS 101 was shown to contain the lowest GC content of IS elements, contrasting to the two hot spring strains *Synechococcus* sp. JA-3-3Ab and *T. elongatus* BP-1 with GC contents of ISs reaching 60% and 53% respectively.

3.2. Subfamily—a lower classification unit of IS elements

132 IS subfamilies were identified in the cyanobacterial genomes in the present study. Among them, ten subfamilies containing the ORF coding region with high homologous to transposase annotated in GenBank cannot match any homologies in the IS Finder, and thus are marked as 'Undefined' (Additional File 2). The copy number of the IS elements in one subfamily ranged from two to ninety-seven (048M843 subfamily). One subfamily was found to be mostly shared by only six strains within the 17 examined strains, indicating that universe subfamilies hardly exist. The phylogeny based on either the IS nucleotide sequences or transposase amino acid sequences within a subfamily were not well consistent to the 16S rDNA based phylogeny (Fig. 3).

Fifty-five subfamilies were found in the genomes of the two *Microcystis* strains, and thirty of them were shared by both strains, while the remaining sixteen and nine subfamilies were present individually. The thirty shared subfamilies including 361 IS elements in *M. aeruginosa* NIES843 and 259 IS elements in *M. aeruginosa* PCC7806, respectively. The filamentous heterocystous strains *Anabaena* sp.

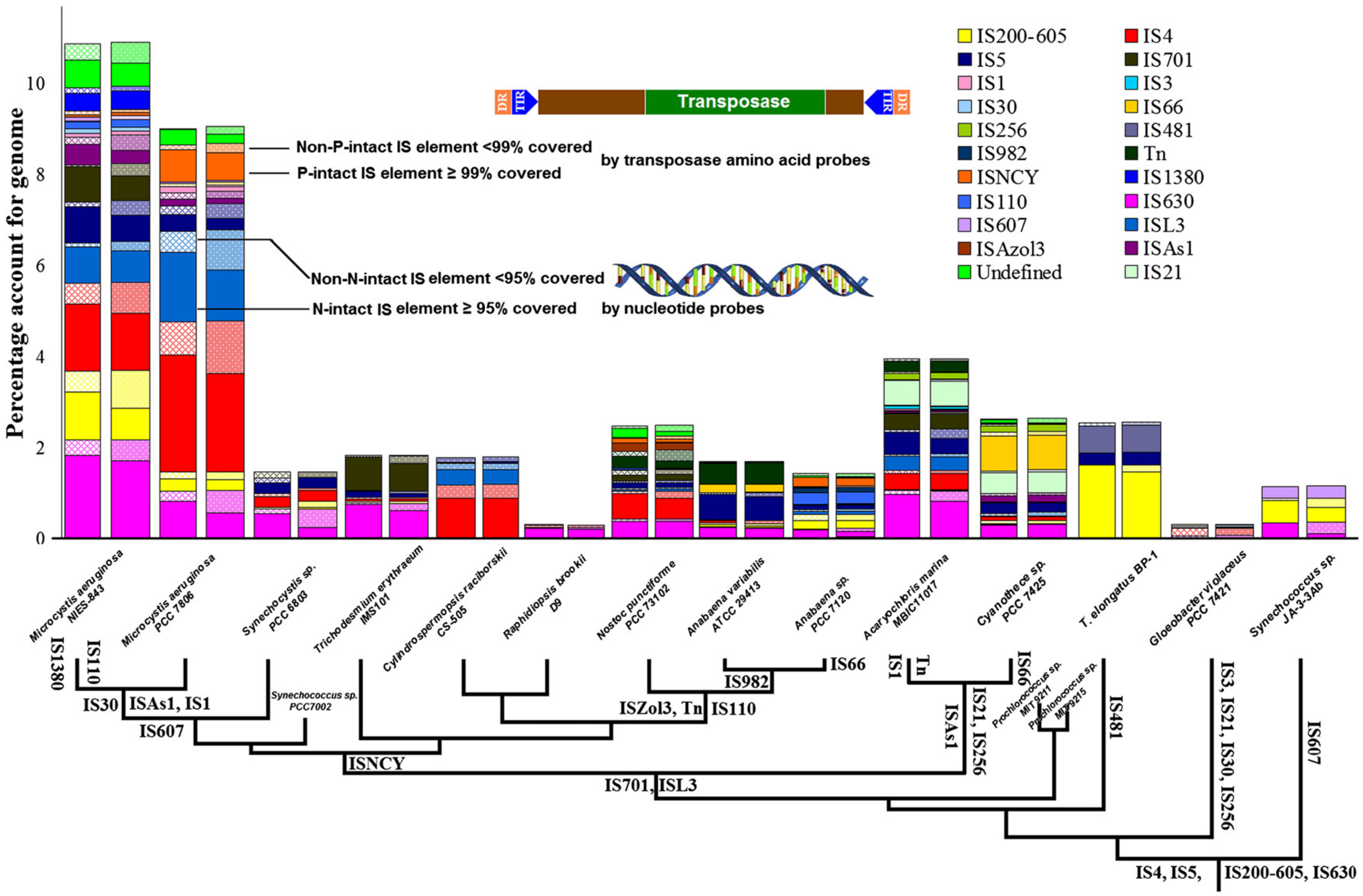

Fig. 1. The IS family composition of seventeen cyanobacterial genomes. For each strain, the left and right columns represent the N-intact and P-intact IS distributions respectively. Grid columns represent non-intact elements. The lower figure is the 16S rDNA sequences based phylogeny of the strains investigated. For each IS family we highlight the most parsimonious scenario of IS families gained by mapping acquisition of elements at each node. The distribution of IS families were also indicated for each strains.

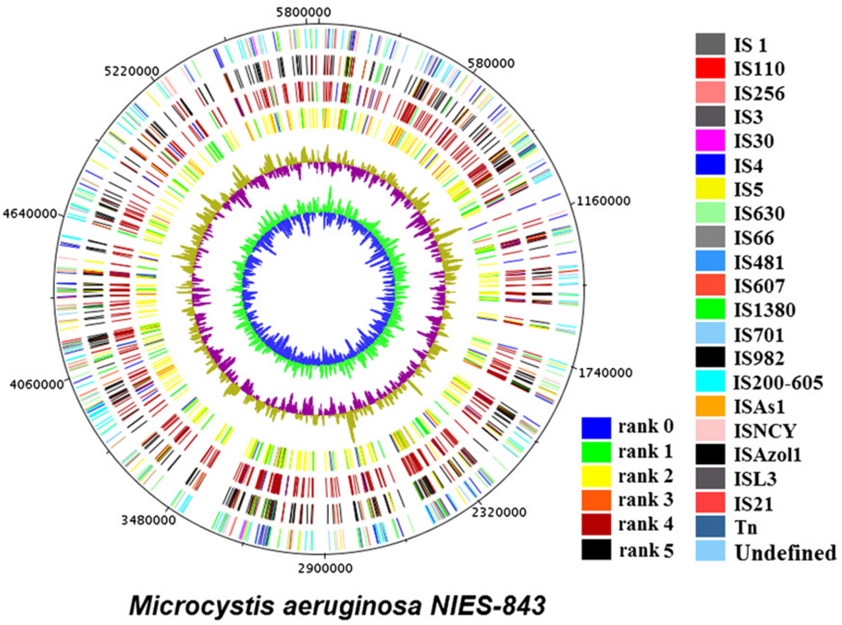

Fig. 2. The insert element map portrayed in the circular chromosome of *Microcystis aeruginosa* NIES 843 genomes. The scale indicates location in bp. The bars marked from outmost circle to the inner ones with colorful marks corresponding to the different IS families, the coverage rank, the similarity rank and the length rank, the GC plot and GC skew respectively. The rank setting for the coverage of transposase amino acid sequence: rank 5: 99%–100%; rank 4: 80%–99%; rank 3: 60%–80%; rank 2: 40–60%; rank 1: 20%–40% and rank 0: <20%. The rank setting for similarity: rank 4: 0.9–1; rank 3: 0.8–0.9; rank 2: 0.7–0.8; rank 1: 0.6–0.7 and rank 0: <0.7. The rank setting for length: rank 4: >3000 bp; rank 3: 2000–3000 bp; rank 2: 1000–2000 bp; rank 1: 500–1000 bp and rank 0: <500 bp.

PCC7120 and *A. variabilis* ATCC 29413 contain thirty-three subfamilies, seven of which were shared by both strains. Twenty-one IS elements from *Anabaena* sp. PCC7120 were shown to have homologous IS elements in *A. variabilis* ATCC29413 genome, and the percentage of homologous elements in two strains is higher than 24%. Compared to the seventy-one of IS elements contained in the hot spring strain of *Synechococcus* sp. JA-3-3Ab, only one IS was found in the plasmid of the freshwater strain *Synechococcus* sp PCC7002. It is seemingly shown that the cyanobacterial strains isolated from hot spring have less IS subfamilies, since only six and four were respectively found in *Synechococcus* sp. JA-3-3Ab and *T. elongatus* BP-1.

3.3. IS family composition in cyanobacterial genomes

93% of the predicted IS elements could be classified into twenty-one bacterial IS families (Fig. 1). Compared with the IS elements in archaea, six IS families including IS3, IS1380, IS701, ISAs1, ISNCY and Tn, were only found in cyanobacteria, while ISA1214, ISM1, IS1595, ISBst12, IS1182, ISH6 and ISC1217 were not found with any homologues in cyanobacteria. IS4, IS5, IS630 and IS200-605 were four dominant and widely distributed IS families in these cyanobacterial genomes. *M. aeruginosa* NIES843 and *A. marina* MBIC11017 contained thirteen IS families, while the two hot spring strains were shown to have only three IS families. It is apparently shown that IS discrepancies exist among the morphologically similar strains. For instance, IS families including IS701, IS30, IS110 and IS1380 detected in *M. aeruginosa* NIES843 were not found any homologous ones in *M. aeruginosa* PCC7806, while nine of fourteen IS families were shared by the both *M. aeruginosa* strains.

3.4. Estimated ancestral IS families

3.4.1. IS4 family

333 IS elements contained by eight cyanobacterial strains were included in IS4 family. And these IS elements could be further classified into twenty IS subfamilies. The phylogenetic relationship among the twenty subfamilies was constructed in this study. As shown in Fig. 4, all these subfamilies were shown to be significantly divided into four clusters. Most of the IS elements within the same IS groups defined by IS Finder could be included in a cluster, such as IS elements from group 10, group 50 and group IS4 Sa. However, two IS elements of group 1634 in IS Finder were separated into cluster III and cluster IV, though these two clusters were closely related in the phylogenetic tree.

3.4.2. IS5 family

IS5 family contained 223 IS elements from eight cyanobacterial strains, and all these IS elements could be further classified into fourteen IS subfamilies. The phylogenetic relationship among these subfamilies showed that thirteen of them, together with eleven records of IS sequences from IS Finder, could be divided into three dominant clusters (Fig. 4). The IS elements from group 1031 and group 903 were located in cluster I and cluster II respectively, with an exception by one IS sequence from group IS427. The IS elements from group ISL2 and group IS5 were gathered in cluster. The cluster III could be further divided into two sub-clusters: the two IS sequences from group ISL2 and one IS sequence from group IS5 were in sub-cluster IIIa, while the other three IS sequences from group IS5 in sub-cluster IIIb.

3.4.3. IS630 family

The IS elements identified as IS630 family could be found in eleven cyanobacterial strains. 430 IS elements belonging to thirty IS subfamilies showed an extremely high level of internal divergences in this IS family. The phylogenetic relationship among these IS subfamilies was constructed. Twenty three of IS subfamilies were divided into five dominant clusters, while the others formed dispersed lineage (Fig. 4).

3.4.4. IS200-605 family

In IS200-605 family, 217 IS elements from ten cyanobacterial strains were included and were further classified into ten IS subfamilies. The phylogenetic relationship among these ten IS subfamilies in IS200-605 family showed that all of these subfamilies could be divided into two dominant clusters (Fig. 4). Four IS elements of group 1341 and two IS elements of group 200 were gathered in cluster I and cluster II respectively.

Table 2
The IS and MITE elements distributing in the cyanobacterial genomes.

	IS											MITE						
Cyanobacteria strains	Genome size	IS frequency	All IS percentage %	P-Intact IS	P-Intact IS percentage %	N-Intact IS	N-Intact IS percentage %	Average length	Min length	Max length	Subfamily number	Type I MITEs	Type II MITEs	MITEs all	All MITE percentage %	MITE GC%	IS GC%	Genome GC%
Microcystis aeruginosa NIES-843	5,842,795	534	10.85	309	7.02	375	8.66	1187	188	2451	47	1110	1356	2466	8.76	39.2	38.6	42.0
Microcystis aeruginosa PCC 7806	5,172,804	359	8.98	186	5.34	240	6.93	1294	285	3696	39	890	1133	2023	8.16	36.2	36.4	42.0
Synechocystis sp. PCC 6803	3,573,470	58	1.43	24	0.66	38	1.03	878	350	1175	8	113	98	211	1.29	39.7	37.2	47.0
Anabaena sp. PCC 7120	6,413,771	56	0.98	43	0.77	46	1	1121	492	1525	15	47	133	180	0.65	43.8	41.1	41.0
Plasmid 7120alpha	408,101	23	6.66	14	4.62	16	5.18	1183	643	1677	9	24	3	27	1.72	36.6	38.6	40.5
Plasmid 7120beta	18,614	3	14.12	0	0.00	0	0	876	553	1049	3	0	0	0	0	0	41.1	40.2
Plasmid 7120gamma	101,965	4	4.35	1	1.34	2	3	1108	670	1364	4	0	3	3	0.75	34.3	42.9	41.0
Plasmid 7120zeta	5,584	0	0	0	0.00	0	0	0	0	0	0	0	0	0	0	0	0	44.2
Plasmid 7120delta	55,414	0	0	0	0.00	0	0	0	0	0	0	0	0	0	0	0	0	41.6
Plasmid 7120epsilon	40,340	0	0	0	0.00	0	0	0	0	0	0	0	0	0	0	0	0	40.9
Gloeobacter violaceus PCC 7421	4,659,019	16	0.31	0	0.00	0	0.00	889	587	1089	6	4	38	42	0.18	59.7	52.1	61.0
Acaryochloris marina MBIC11017	6,503,724	188	3.47	141	3	164	3.19	1200	315	4584	30	214	274	488	1.75	49.1	0	47.0
Plasmid AcarypREB1	374,161	4	2.23	3	1.96	3	1.96	2083	1349	4584	4	0	0	0	0	0	0	47.3
Plasmid AcarypREB2	356,087	16	7.63	13	6.93	15	7.30	1698	775	4584	11	0	6	6	0.61	45.8	0	45.3
Plasmid AcarypREB3	273,121	16	6.25	8	4.01	10	4.03	1067	493	2670	12	0	0	0	0	0	0	45.2
Plasmid AcarypREB4	226,680	5	3.52	5	3.52	5	3.52	1598	1060	2669	5	0	0	0	0	0	0	45.9
Plasmid AcarypREB5	177,162	6	4.86	6	4.86	5	3.56	1435	1060	2297	6	0	0	0	0	0	0	44.7
Plasmid AcarypREB6	172,728	9	7.87	4	5.56	4	5.56	1510	481	4603	5	0	12	12	1.96	47.4	0	47.1
Plasmid AcarypREB7	155,110	3	3.38	2	2.62	2	2.62	1749	1183	2669	2	0	0	0	0	0	0	45.6
Plasmid AcarypREB8	120,693	5	7.12	2	3.35	3	5.39	1719	864	2669	4	0	3	3	0.3	64.2	0	45.4
Plasmid AcarypREB9	2,133	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42.5
Anabaena variabilis ATCC29413	6,365,727	53	1.37	43	1.19	48	1.30	1648	456	6495	11	83	117	200	0.74	44.5	42.0	41.0
Plasmid AnabA	366,354	10	4.5	7	3.90	7	3.90	1649	595	6495	6	18	0	18	1.11	45.1	42.5	40.5
Plasmid AnabB	35,762	0	0	0	0.00	0	0.00	0	0	0	0	0	0	0	0	0		38.5
Plasmid AnabC	300,758	7	4.28	4	3.14	5	3.75	1838	500	6495	6	0	0	0	0	0	40.7	42.0
Nostoc punctiforme PCC 73102	8,234,322	146	2.03	100	1.49	119	1.75	1143	419	4826	27	258	305	563	1.45	39.8	38.9	41.0
Plasmid pNUN01	354,564	14	5.02	6	2.82	8	3.37	1271	548	4826	9	3	0	3	0.22	36.6	37.8	40.5
Plasmid pNUN02	254,918	14	7.72	7	3.7	9	4.33	1405	681	4824	11	0	0	0	0	0	36.6	40.7
Plasmid pNUN03	123,028	4	9.78	0	0	1	3.92	3009	1002	5031	3	0	0	0	0	0	40.5	40.9
Plasmid pNUN04	65,940	2	8.7	1	7.32	1	7.32	2868	908	4828	2	0	0	0	0	0	39.4	41.5
Plasmid pNUN05	26,419	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	42.3
Synechococcus sp. PCC 7002	3,008,047	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	49.0
Plasmid 7002pAQ1	4,809	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	49.0
Plasmid 7002pAQ2	16,103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	45.9
Plasmid 7002pAQ4	31,972	1	1.82	1	1.82	1	1.82	582	582	582	1	0	0	0	0	0	50.0	44.1
Plasmid 7002pAQ5	38,515	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	42.6
Plasmid 7002pAQ6	124,030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	45.1
Plasmid 7002pAQ7	18,459	1	3.15	1	3.15	1	3.15	582	582	582	1	0	0	0	0	0	50.0	47.3
Cyanothece sp. PCC 7425	5,374,574	91	2.09	67	1.73	71	1.8	1233	382	2666	17	95	101	196	1.01	52.9	52.2	50.0
Plasmid 742501	196,837	6	4.57	4	3.69	4	3.69	1500	802	2665	5	3	0	3	0.47	53.2	53.1	48.9
Plasmid 742502	179,973	23	16.2	14	12.65	14	12.65	1268	456	2664	11	14	9	23	3.72	52.9	52.8	49.1
Plasmid 742503	34,726	0	0	0	0	0	0	0	0	0	0	0	0	0	0			47.1
Synechococcus sp. JA-3-3Ab	4,659,019	71	1.14	47	0.68	68	1.10	747	417	1054	6	85	147	232	1.55	54.3	52.1	60.0
Thermosynechococcus elongatus BP-1	2,593,857	58	2.53	52	2.31	55	2.47	1130	348	1473	4	100	138	238	1.93	51.8	49.9	53.0
Trichodesmium erythraeum IMS101	7,750,108	106	1.53	83	1.24	93	1.41	1130	353	1386	12	0	0	0	0	0	34.0	34.0
Cylindrospermopsis raciborskii cs-505	3,879,030	58	0.89	31	1.28	31	1.31	1227	281	2202	4	185	631	816	3.86	39.5	32.9	40.2
Raphidiopsis brookii D9	3,186,511	10	0.29	6	0.20	6	0.24	927	504	1105	4	3	7	10	0.09	36.9	42.6	40.1

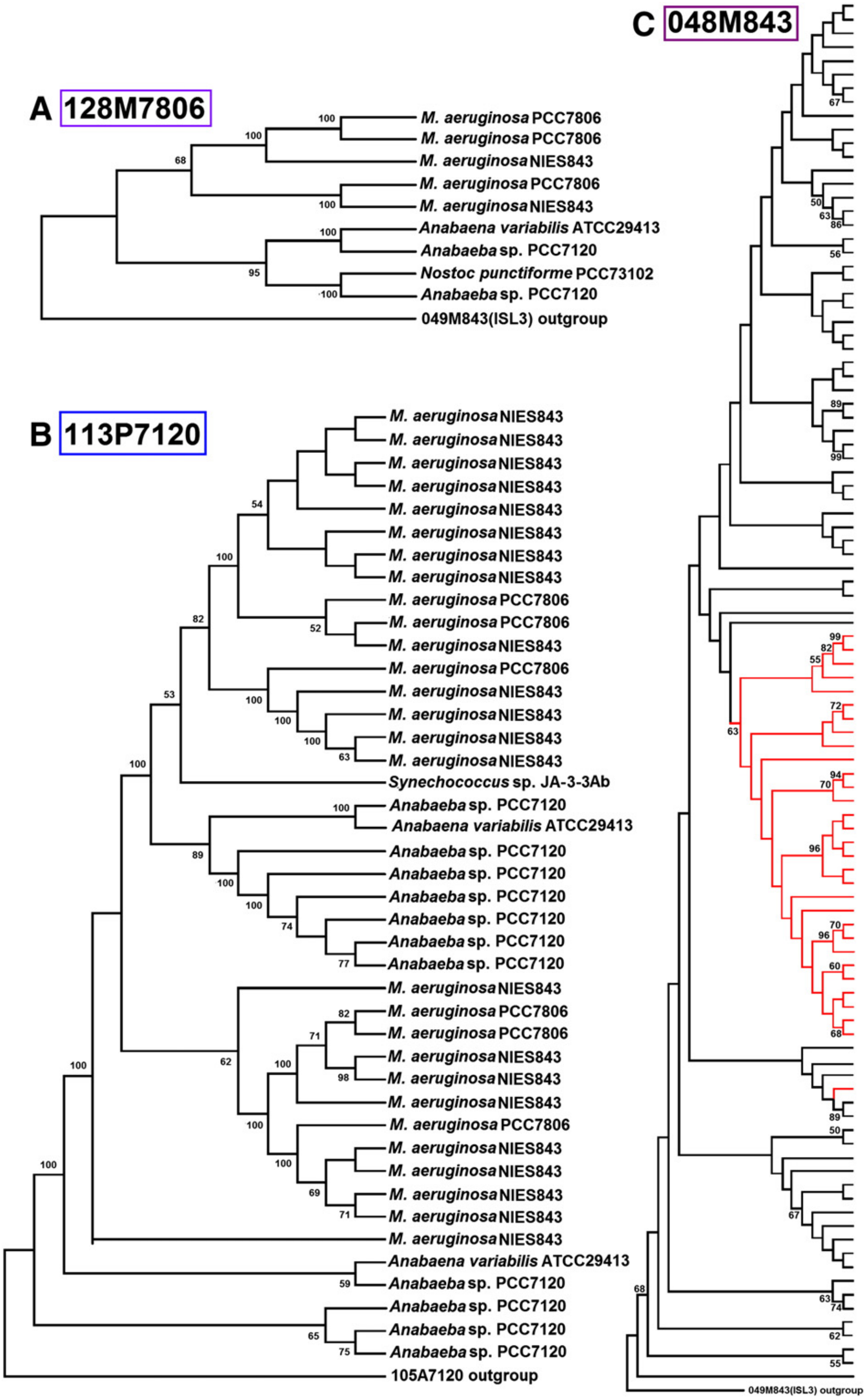

Fig. 3. Phylogenies based on the all the IS nucleotide probe sequences of subfamilies 113P7120, 128M7806 and 048M843. 3A. the phylogeny based on the nucleotide probe sequences of the IS subfamily 128M7806; 3B. the phylogeny based on the nucleotide probe sequences from IS subfamilies 113P7120; 3C. the phylogeny based on the nucleotide probe sequences of the IS subfamily 048M843. All the clades in black represent the clades of ISs from *M. aeruginosa* PCC7806, while the clade lines in red represent the clades of ISs from *M. aeruginosa* NIES843. Bootstrap values greater than 50% with neighbor-joining methods are indicated on the trees.

3.5. The IS intactness diversity

The intactness of transposase ORF is the most important factor in determining the autonomous transposable action. Segment loss, nucleotide mutations, insertions, and deletions caused by reading frame interrupted or shift are the principal mechanisms for interrupting the intactness. The number of P-intact IS elements in the examined cyanobacterial genomes was 1240, accounting for 62.6% of all the predicted IS elements. 74.0% of these P-intact sequences were further found to have more than 99% similarities with the probe sequences. The IS elements shorter than 500 bp were mostly considered to be non-P-intact. The percentages of the P-intactness in different IS families were different, from 50% (Tn family) to 100% (IS982 family). *M. aeruginosa* NIES 843 was found to contain 10% higher abundance of the P-intact IS

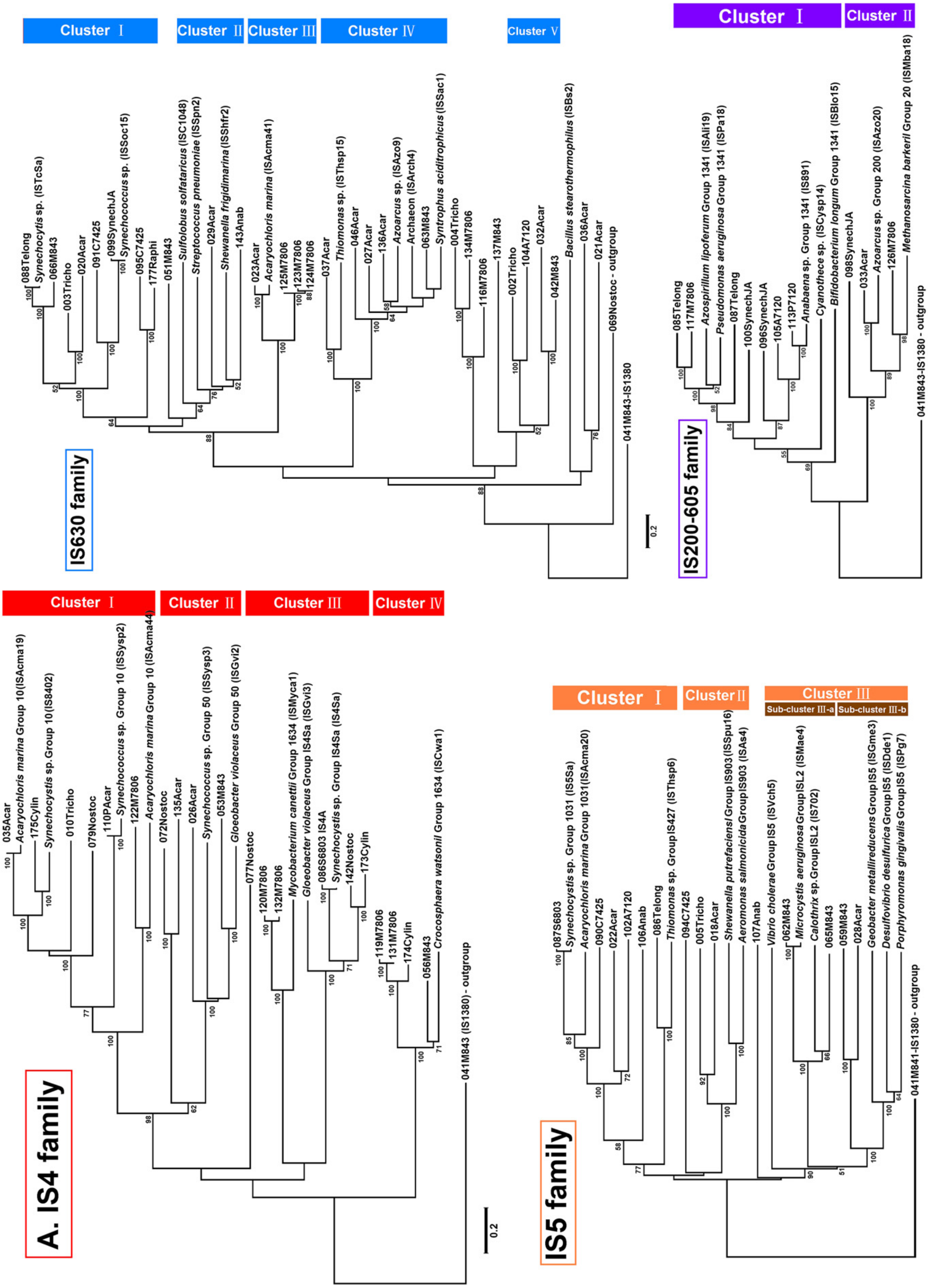
A. IS4 family
Cluster Ⅰ
Cluster Ⅱ
Cluster Ⅲ
Cluster Ⅳ
035Acar
Acaryochloris marina Group 10(ISAcma19)
175Cylin
Synechocystis sp. Group 10(IS8402)
010Tricho
079Nostoc
110PAcar
Synechococcus sp. Group 10 (ISSysp2)
122M7806
Acaryochloris marina Group 10 (ISAcma44)
072Nostoc
135Acar
026Acar
Synechococcus sp. Group 50 (ISSysp3)
053M843
Gloeobacter violaceus Group 50 (ISGvi2)
077Nostoc
120M7806
132M7806
Mycobacterium canettii Group 1634 (ISMyca1)
Gloeobacter violaceus Group IS4Sa (ISGvi3)
086S6803 IS4A
Synechocystis sp. Group IS4Sa (IS4Sa)
142Nostoc
173Cylin
119M7806
131M7806
174Cylin
056M843
Crocosphaera watsonii Group 1634 (ISCwa1)
041M843 (IS1380) - outgroup
0.2
IS630 family
Cluster Ⅴ
088Telong
Synechocystis sp. (ISTcSa)
066M843
003Tricho
020Acar
091C7425
099SynechJA
Synechococcus sp. (ISSoc15)
095C7425
177Raphi
051M843
Sulfolobus solfataricus (ISC1048)
Streptococcus pneumoniae (ISSpn2)
029Acar
Shewanella frigidimarina (ISShfr2)
143Anab
023Acar
Acaryochloris marina (ISAcma41)
125M7806
123M7806
124M7806
037Acar
Thiomonas sp. (ISThsp15)
046Acar
027Acar
136Acar
Azoarcus sp. (ISAzo9)
Archaeon (ISArch4)
063M843
Syntrophus aciditrophicus (ISSac1)
004Tricho
134M7806
116M7806
137M843
002Tricho
104A7120
032Acar
042M843
Bacillus stearothermophilus (ISBs2)
036Acar
021Acar
069Nostoc - outgroup
041M843-IS1380
IS5 family
Sub-cluster Ⅲ-a
Sub-cluster Ⅲ-b
087S6803
Synechocystis sp. Group 1031 (IS5Sa)
Acaryochloris marina Group 1031(ISAcma20)
090C7425
022Acar
102A7120
106Anab
086Telong
Thiomonas sp. Group IS427 (ISThsp6)
094C7425
005Tricho
018Acar
Shewanella putrefaciens Group IS903 (ISSpu16)
Aeromonas salmonicida Group IS903 (ISAs4)
107Anab
Vibrio cholerae Group IS5 (ISVch5)
062M843
Microcystis aeruginosa Group ISL2 (ISMae4)
Calothrix sp. Group ISL2 (IS702)
065M843
059M843
028Acar
Geobacter metallireducens Group IS5 (ISGme3)
Desulfovibrio desulfurica Group IS5 (ISDde1)
Porphyromonas gingivalis Group IS5 (ISPg7)
041M841-IS1380 - outgroup
0.1
IS200-605 family
085Telong
117M7806
Azospirillum lipoferum Group 1341 (ISAli19)
Pseudomonas aeruginosa Group 1341 (ISPa18)
087Telong
100SynechJA
096SynechJA
105A7120
113P7120
Anabaena sp. Group 1341 (IS891)
Cyanothece sp. (ISCysp14)
Bifidobacterium longum Group 1341 (ISBlo15)
098SynechJA
033Acar
Azoarcus sp. Group 200 (ISAzo20)
126M7806
Methanosarcina barkeri Group 20 (ISMba18)
041M843-IS1380 - outgroup

Fig. 4. Phylogenies based on transposase amino acid sequences of the putative ancestral IS families in cyanobacteria. Bootstrap values greater than 50% with neighbor-joining methods are indicated on the trees. The records with brackets were from IS Finder database.

elements than *M. aeruginosa* PCC7806. Subfamily 048M843 contained the highest abundance of IS element copy. Sixty-three IS elements in this subfamily detected in the genomes of *M. aeruginosa* NIES843 and *M. aeruginosa* PCC7806 were P-intact ones, while four pieces of IS elements in *M. aeruginosa* NIES 843 and one in *M. aeruginosa* PCC7806 sharing the same nucleotide substitution were ORF-fractured ones.

N-intact IS elements were shown to be partly different from the P-intact ones. More than 98.3% of the P-intact IS elements were simultaneously defined as N-intact IS elements, and 82.7% N-intact IS elements are composed by the P-intact IS elements. The average percentage of the N-intact IS elements is 74.8%, ranging from 62.1% to 100%. The percentage of the N-intact IS in the genomes of the two hot spring strains was high, reaching 94.8% and 95.7%, respectively. Neither N-intact nor P-intact IS could be detected in the genome of *Gloeobacter violaceus* PCC7421.

3.6. Nucleotide and protein sequence diversity in IS elements

The phylogenetic analysis based on all of the IS nucleotide sequences within subfamilies 113P7120, 128M7806 and 048M843, which are representatives of the most extensive strain resources, highest subfamily divergence and most copy number, was executed respectively. In subfamily 048M843, the nucleotide sequence divergence of the IS elements from *M. aeruginosa* PCC 7806 was much higher than that from *M. aeruginosa* NIES843 (Fig. 3). The IS elements from *M. aeruginosa* NIES843 were mostly gathered in one lineage, further reflecting that the ORF-fractured segments were mixed with the intact ones. The only one ORF-fractured IS element from *M. aeruginosa* NIES843 was clustered together with the IS elements from *M. aeruginosa* PCC7806. In subfamily 128M7806, *M. aeruginosa* PCC 7806 and *M. aeruginosa* NIES 843 are distantly separated from two *Anabaena* strains. In subfamily 113P7120, the IS elements were mainly from two *Microcystis* strains and two *Anabaena* strains. The phylogeny based on the IS nucleotide sequences showed that the IS elements from *Microcystis* form four clusters, while the IS elements from *Anabaena* were grouped as two clusters. It is shown that one genome may contain many IS elements of one subfamily from extensive resources. The IS elements from *Cyanothece* sp. PCC 7425 and *Synechococcus* sp. JA-3-3Ab form a single cluster away from others.

Diversity index of both nucleotide and transposase amino acid sequences from the P-intact IS elements of the 132 subfamilies were calculated (Supplemental Table S1). The highest nucleotide and amino acid divergences were found in the subfamily 128M7806, with the index values as 0.21656 and 0.9289 respectively. High conservation of transposase amino acid sequences in 42 IS subfamilies was also shown, with their protein diversity indices as 0. Twelve subfamilies with high conservation of protein sequence correspond to vary of nucleotide sequences.

3.7. MITE in cyanobacterial genomes

Totally 7763 MITEs were identified in these cyanobacterial genomes, and 3249 pieces of them can be classified as type I. All the type I MITEs detected in this study have been found to be IS originated. The remaining 4514 MITE elements were classified as type II. The length of most MITEs ranged from 100 bp to 499 bp (Supplemental Fig. S2). The abundance is inversely correlated to the length of MITEs, and 60% of MITEs were in the length ranging between 120 and 260 bp. The frequency of the MITEs in cyanobacterial genomes analyzed in this study varied from 0 to 2466 pieces, taking the percentages from 0 to 8.76%. The highly linear correlation between the IS and MITE elements was found in this study. The correction coefficients for the frequency of IS vs type I MITE, IS vs type II MITE and IS vs all MITE reach 92.3%, 81.8% and 87.5% respectively (Supplemental Fig. S3). The frequency of type II MITEs was one to three times higher than that of type I ones, with the exception for the genomes of *Synechocystis* sp. PCC6803 and two plasmids from the strains PCC 7120 and PCC7425. Unexpectedly, the TIR border couldn't be detected in the genome of *T. erythraeum* IMS101. Similar to IS elements, MITEs have no AT or GC bias. The lowest GC content of IS elements was 36.2% in *M. aeruginosa* PCC 7806 genome and the higher ones were found in *Synechococcus* sp. JA-3-3Ab and *T. elongatus* BP-1 inhabiting in hot spring, the percentage of which were 60% and 53% respectively.

4. Discussion and conclusions

Cyanobacteria have been considered to originate about 2.7 billion years ago (Timothy, 2007), and information on cyanobacterial transposable elements in such a long term would certainly help to understand their roles along the evolutionary course. This study demonstrated an extremely high and hierarchical diversity of transposable elements in cyanobacterial phylum.

The big difference in the abundance of transposable element system was found among cyanobacterial genomes. Zhou et al. (2008) assumed that the frequency of recently active IS elements, which are similar to the defined P-intact elements in this study, positively correlate with genome size (Zhou et al., 2008). However, the analysis on the transposable element system from recently released cyanobacterial genomes revealed that the frequencies of IS, P-intact and N-intact IS elements have no significant relationship with the genome size (Supplementary Fig. S5). The highest abundance of transposable elements was found in the unicellular *M. aeruginosa* strains with the medium size of genome, while the filamentous *A. variabilis* ATCC29413 and *N. punctiforme* PCC 73102 strains with genome size larger than 6 Mbp were revealed to have smaller and simpler transposable element systems. Genome plasticity in prokaryotes is often considered to be an adaptive strategy allowing microorganisms to promote diversification in the way similar to sexual reproduction in eukaryotic organisms (Filée et al., 2007). Frangeul et al. (2008) pointed that a high frequency of transposable elements inhabiting in genomes would facilitate this adaptive strategy (Frangeul et al., 2008). High abundance of transposable elements found in the *M. aeruginosa* strains examined here demonstrate that their genomes may be rearranged to cause positive mutations accelerating adaptations to various freshwater ecosystems, and this high genome plasticity caused by genomic rearrangement might be an explanation to the fact that *Microcystis* is the most successful organism to compete over others. *Microcystis* species have been globally found as the dominant species, to largely grow in eutrophic freshwaters. *M. aeruginosa* NIES843 and *M. aeruginosa* PCC7806 strains were respectively isolated from Lake Kasumigaura of Japan in 1997 and from Braakman reservoir of Netherlands in 1972, and the difference of IS composition and abundance between the two strains may be caused by the different habitant environment and strain maintenance periods.

IS family and subfamily are two hierarchical classification levels for cyanobacterial transposable element systems. In contrast to the lower classification unit 'IS group' raised by IS Finder database, IS subfamily as the basic classification unit in transposable element system is firstly proposed in this study. As shown in Fig. 4, most of the IS sequences assigned a 'group' label could be orderly clustered. However, after appending more and more new identified IS elements from newly sequenced species, the phylogenetic relationship would be gradually adjusted, therefore causing some groups to be relabeled. In some IS family of hypervariabilty such as IS630 (Fig. 4), the 'cluster' gathering within a group was difficult to obtain, thus the 'group' label was hardly assigned. Nucleotide probe library is a necessary component for transposable element mining. The definition of IS subfamily base on a stable and persistently renewable IS probe library, instead of instable phylogeny related was assumed to be an easy-defined and reliable unit in IS element system classification. The divergence of both IS family and subfamily composition and their nucleotide and transposase amino acid sequences shown in this study also reflected the hypervariabilty of the transposable elements in cyanobacterial genomes. 21 IS families and 132 subfamilies were identified in cyanobacteria genomes examined here. Based on the widely confirmed 16S rRNA phylogeny and the IS family composition for each strains, we dedicate the most parsimonious evolutionary scenario of IS acquisition for each family (Fig. 1). Santiago et al. (2002) indicated that in *Arapdopsis*, the more variable a transposable element family (subfamily) is, the more ancient the amplification burst that has generated it should be (Santiago et al., 2002). Similarly, four IS families in this study, IS4, IS5, IS 605 and IS630, which were found to exhibit a wide distribution and diversity in cyanobacterial genomes. 1203 IS elements from these four IS families accounts for 60.7% of all the IS elements and each IS family was shared by more than eight species. Therefore, these four could be considered as cyanobacterial ancestral IS families. The phylogeny based on the nucleotide sequences of the widely distributed IS subfamilies revealed that the IS elements from one genome commonly gathered together and the IS elements from close related species have high similarity of nucleotide sequences than that between distantly related species (Fig. 3). Such a result implied that the most likely exchange and replication of the transposable elements in cyanobacteria may occur within a genome, followed by close related species. Furthermore, more resources of IS elements belonging to one IS family were also found in one genome, which may provide valuable information to analyze the population relationship and species evolution in the future.

In eukaryotes, recent transposable element insertions have been used in population genetics studies and regarded as identical-by-descent genetic markers for the evolution, forensics and population history studies [(Lozano et al., 2010), (Engel et al., 2001), (Hammer, 1994) and (González et al., 2008)]. A transposable element family/subfamily insertion with lower nucleotide divergence (<1% or lower) has been considered as a recent insertion [(Lozano et al., 2010) and (González et al., 2008)]. Among all the IS subfamilies examined in the cyanobacterial genomes, many of them were shown to have a lower nucleotide diversity (Additional File), and thirty IS subfamilies even having the nucleotide diversity index as zero. Therefore, these IS subfamilies with lower diversity index were considered as the putative recent IS subfamily insertions, which have the potential used for the analyses of cyanobacterial population relationship in the future.

In most of the examined cyanobacterial genomes, the intact IS elements showed to contain more copies and higher sequence diversity than the fractured ones. Surprisingly, *G. violaceus* PCC7421 was the only strain without the intact IS elements, which cannot be explained so far. Many ORF-fractured transposase still showed to have the basic structure of the N-intact elements, but the fracture of these transposases may attribute to the fact that their coding frames are interrupted by slipped strand mispairing during DNA replication on a single DNA strand, as described by Bichara et al. (2006).

Previous studies indicated that unique morphological, physiological and genetic characters were always found in organisms from the extreme environments [(Badyaev and Foresman, 2000) and (Rothschild and Mancinelli, 2001)]. Zhou et al. (2008) concluded that hot spring seems to be one of the favorite living environments for organisms with active IS elements (Zhou et al., 2008). In the present study, a medium content of IS elements contained in *Synechococcus* sp. JA-3-3Ab and *T. elongatus* BP-1 inhabiting in hot spring environments are revealed to have higher intactness of IS family and subfamily compositions. Such results suggest that a high percentage of intact IS might play a partial role in maintaining the genome stability in the extreme environments.

Although MITE element system was described in the genome of *M. aeruginosa* NIES 843 (Kaneko et al., 2007), the information about MITE in prokaryotes is still scarce. In this study, higher abundance of MITEs and two types of MITEs revealed in cyanobacterial genomes provided a basic overview for the knowledge of MITEs in cyanobacteria. Actually, type I MITE was assumed to be a result of a deletion within an IS element and called as 'parasites of parasites' as well [(Brügger et al., 2002) and (González and Petrov, 2009)], thus many of non-intact IS elements belonged to the type I MITE. However, it is still hard to implicate cyanobacterial MITEs as the diversity indicator since they are too short and irregular.

Conclusively, the analyses on the transposable system of cyanobacterial genomes will help to improve understanding the knowledge for the diversity of cyanobacteria. The features of the transposable elements in cyanobacteria, including the abundance of intact IS, the composition of IS families and subfamilies, the sequence diversity of IS element nucleotide and transposase amino acid, have shown to be valuable indicators for studies on cyanobacterial diversity. It is specially noted here that the *Microcystis* strains contain a high abundance of IS elements, which allows us to use the transposable element system as a new perspective to further explore the diversity and population relationship of water bloom forming cyanobacterial species.

Acknowledgements

We thank Dr. Fengfeng Zhou (UGA, US), Prof. Mick Chandler (C.N.R.S, France) and the anonymous reviewers for the valuable discussion, suggestions and arguments. This research is funded by the National Key Basic Research Program (973) (2008CB418002), FEBL fund (2011FB17) and the CAS-MPG joint doctoral program.

Author's contributions

SL, RL and SH designed this study. SL and PX performed the data mining and analysis. TZ and SH made important and meaningful comments; SL and RL wrote this manuscript. MV provided this program a powerful platform. All authors read and approved the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.gene.2010.11.011.

References

Badyaev, A.V., Foresman, K.R., 2000. Extreme environmental change and evolution: stress-induced morphological variation is strongly concordant with patterns of evolutionary divergence in shrew mandibles. P. Roy. Soc. B Biol. Sci. 267, 371–377.

Barnes, M.J., Lobo, N.F., Coulibaly, M.B., Sagnon, N., Costantini, C., Sansky, N.J., 2005. SINE insertion polymorphism on the X chromosome differentiates *Anopheles gambiae* molecular forms. Insect Mol. Biol. 14, 353–363.

Bichara, M., Wagner, J., Lambert, I.B., 2006. Mechanisms of tandem repeat instability in bacteria. Mutation research/fundamental and molecular mechanisms of mutagenesis. Mutat. Res Fund. Mol. M. 598, 144–163.

Boulesteix, M., Simard, F., Antonio-Nkondjio, C., Awono-Ambene, H.P., Fontenille, D., Biémont, C., 2007. Insertion polymorphism of transposable elements and population structure of *Anopheles gambiae* M and S molecular forms in Cameroon. Mol. Ecol. 16, 441–452.

Brügger, K., et al., 2002. Mobile elements in archaeal genomes. FEMS Microbiol. Lett. 206, 131–141.

Chen, Y., Zhou, F., Li, G., Xu, Y., 2009. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. Gene 436, 1–7.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Engel, A.M.R., et al., 2001. Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159, 279–290.

Filée, J., Siguier, P., Chandler, M., 2007. Insertion sequence diversity in archaea. MMBR 71, 121–157.

Frangeul, L., et al., 2008. Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. BMC Genomics 9, 274.

González, J., Petrov, D., 2009. MITEs—The ultimate parasites. Science 325, 1352–1353.

González, J., Lenkov, K., Lipatov, M., Macpherson, J.M., Petrov, D.A., 2008. High rate of recent transposable element–induced adaptation in *Drosophila melanogaster*. PLoS Biol. 6, e251.

Gray, Y., 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. 16, 461–468.

Hammer, M.F., 1994. A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. Mol. Biol. Evol. 11, 749–761.

Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.

Kaneko, T., 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., et al., 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8, 205–213.

Kaneko, T., et al., 2007. Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. DNA Res. 14, 247–256.

Kidwell, M.G., 1992. Horizontal transfer of P elements and other short inverted repeat transposons. Genetica 86, 275–286.

Kidwell, M.G., Lisch, D.R., 2001. Perspective: transposable elements, parasitic DNA and genome evolution. Evolution 55, 1–24.

Kosier, B., Pühler, A., Simon, R., 1993. Monitoring the diversity of *Rhizobium meliloti* field and microcosm isolates with a novel rapid genotyping method using insertion elements. Mol. Ecol. 2, 35–46.

Kurtz, S., 1999. The Vmatch large scale sequence analysis software. Computer program.

Langdon, T., Jenkins, G., Hasterok, R., Jones, R.N., King, P., 2003. A high-copy-number CACTA family transposon in temperate grasses and cereals. Genetics 163, 1097–1108.

Larkin, M.A., et al., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Leavis, H.L., et al., 2007. Insertion sequence–driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. PLoS Pathog. 3, e7.

Lepetit, D., Brehm, A., Fouillet, P., Biémont, C., 2002. Insertion polymorphism of retrotransposable elements in populations of the insular, endemic species *Drosophila madeirensis*. Mol. Ecol. 11, 347–354.

Lozano, L., et al., 2010. Evolutionary dynamics of insertion sequences in relation to the evolutionary histories of the chromosome and symbiotic plasmid of Rhizobium etli populations. Appl. Environ. Microbiol. 76, 6504–6513.

Mulkidjanian, A.Y., et al., 2006. The cyanobacterial genome core and the origin of photosynthesis. Proc. Natl Acad. Sci. USA 103, 13126–13131.

Nakamura, Y., et al., 2002. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. DNA Res. 9, 123–130.

Nakamura, Y., et al., 2003. Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. DNA Res. 10, 137–145.

Nekrutenko, A., Li, W.H., 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 17, 619–621.

Niemann, S., Puhler, A., Tichy, H.V., Simon, R., Selbitschka, W., 1997. Evaluation of the resolving power of three different DNA fingerprinting methods to discriminate among isolates of a natural *Rhizobium meliloti* population. J. Appl. Microbiol. 82, 477–484.

Rothschild, L.J., Mancinelli, R.L., 2001. Life in extreme environments. Nature 409, 1092–1101.

Santiago, N., Herráiz, C., Goñi, J.R., Messeguer, X., Casacuberta, J.M., 2002. Genome-wide analysis of the emigrant family of MITEs of *Arabidopsis thaliana*. Mol. Biol. Evol. 19, 2285–2293.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., Chandler, M., 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34, 32–36.

Smit, A.F.A., Hubley, R., Green, P., 1996–2004. RepeatMasker Open-3.0. http://www.repeatmasker.org1996–2004.

Stucken, K., et al., 2010. The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. PLoS ONE 5, e9235.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Timothy, W.L., 2007. Palaeoclimate: Oxygen's rise reduced. Nature 448, 1005–1006.

Wicker, T., et al., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982.

Wilson, K.M., Schembri, M.A., Baker, P.D., Saint, C.P., 2000. Molecular characterization of the toxic cyanobacterium *Cylindrospermopsis raciborskii* and design of a species-specific PCR. Appl. Environ. Microbiol. 66, 332–338.

Zampicinini, G., Blinov, A., Cervella, P., Guryev, V., Sella, G., 2004. Insertional polymorphism of a non-LTR mobile element (NLRCth1) in European populations of *Chironomus riparius* (Diptera, Chironomidae) as detected by transposon insertion display. Genome 47, 1154–1163.

Zhou, F., Olman, V., Xu, Y., 2008. Insertion sequences show diverse recent activities in Cyanobacteria and Archaea. BMC Genomics 9, 36.