Authors: Akira Hikosaka, Akira Kawahara

Title: A systematic search and classification of T2 family miniature inverted-repeat transposable elements (MITEs) in *Xenopus tropicalis* suggests the existence of recently active MITE subfamilies

The affiliation and address of the authors: Graduate School of Integrated Arts and Sciences, Hiroshima University, 739-8521, Higashi-Hiroshima, Hiroshima, Japan

Corresponding author : Akira Hikosaka

email: akirahs@hiroshima-u.ac.jp

tel: +81-82-424-6567

fax: +81-82-424-0759

Abstract

To reveal the genome-wide aspects of *Xenopus* T2 family miniature inverted-repeat transposable elements (MITEs), we performed a systematic search and classification of MITEs by a newly developed procedure. A terminal sequence motif (T2-motif: TTAAAGGRR) was retrieved from the *Xenopus tropicalis* genome database. We then selected 51- to 1000-bp MITE candidates framed by an inverted pair of 2 T2-motifs. The 34,398 candidates were classified into possible clusters by a novel terminal sequence (TS)-clustering method on the basis of differences in their short terminal sequences. Finally, 19,242 MITEs were classified into 16 major MITE subfamilies (TS subfamilies), 10 of which showed apparent homologies to known T2 MITE subfamilies, and the rest were novel TS subfamilies. Intra- and inter-subfamily similarities or differences were investigated by analyses of diversity in GC content, total length, and sequence alignments. Furthermore, genome-wide conservation of the inverted pair structure of subfamily-specific TS stretches and their target site sequence (TTAA) were analyzed. The results suggested that some TS subfamilies might include active or at least recently active MITEs for transposition and/or amplification, but some others might have lost such activities a long time ago. The present methodology was efficient in identifying and classifying MITEs, thereby providing information on the evolutionary dynamics of MITEs.

Key words: transposon, miniature inverted-repeat transposable element (MITE), T2 family, *Xenopus*

Introduction

Miniature inverted-repeat transposable elements (MITEs) are a subclass of transposable elements first identified in plants (Bureau and Wessler 1992) and later in various eukaryotes, including animals and fungi (for a review, see Feschotte et al. 2002); they have also been reported in a few archaeal (Brügger et al. 2002) and bacterial (Zhou et al. 2008) genomes. MITEs represent nonautonomous, usually short (less than 1kb) class II (DNA-mediated) transposons. They are characterized by their terminal inverted repeat (TIR) structure, short direct repeats formed by target site duplication (TSD), AT richness, and absence of gene-coding capacity. Transposases of class II transposons bind to MITE TIRs (Feschotte et al. 2005; Loot et al. 2006) and mobilize them (Dufresne et al. 2007; Miskey et al. 2007; Yang et al. 2007). Thus, they are assumed to be mobilized via a "cut-and-paste" system with the help of specific transposases encoded by autonomous DNA transposons. Furthermore, MITEs are discriminable from other nonautonomous class II transposons by their high copy numbers (in the order of up to 10^4) and the high uniformity of their copies. These features suggested that MITEs could be rapidly and extensively amplified under certain conditions (amplification burst) (Naito et al. 2006). However, the molecular mechanisms and the evolutionary dynamics by which MITEs extensively amplify themselves in the genome remain unknown.

MITEs are thought to drive evolution of their host's genome. In some cases, copies of MITEs have been "domesticated" by their host species. They have been shown to generate microRNAs in primates (Piriyapongsa and Jordan 2007; Piriyapongsa et al. 2007), transcriptional regulatory elements (El Amrani et al. 2002), and the matrix attachment region (Avramova et al. 1998). Another role played by MITEs in the evolution of the host genome is the formation of simple sequence repeats (SSRs). We have reported

a predominant MITE, Xmix, in the African clawed frog *Xenopus*; this MITE has an amplified internal segment that represents a large family of SSRs (Xstir) (Hikosaka et al. 2000; Hikosaka and Kawahara 2004). It is known that SSR is a genetic element essential for the higher order chromosomal structure (Ugarkovic 2005). Xmix belongs to the T2 MITE family (T2-MITE) that has the TTAA target site and contains the AGGRR TIR motif. Several T2-MITE subfamilies have been identified in frogs (Ünsal and Morgan 1995; Hikosaka et al. 2004) and fish (Izsvák et al. 1999). However, the transposase gene responsible for the mobility of T2-MITEs has not been identified. Therefore, it is entirely unknown whether there are active MITEs and whether all the MITE subfamilies have been created by a single transposase in each animal species. Genome-wide searches and the classification and characterization of various T2-MITEs are necessary for elucidating the evolutionary history and current status of T2-MITEs.

There are currently 2 types of search procedures to identify MITEs from genome databases: the computer programs MAK (Yang and Hall 2003) and FINDMITE (Tu 2001). The former depends on a homology search for known MITEs and would not be able to find new MITEs. The latter utilizes TIR base-matching thresholds to find MITEs, which would restrict its capacity to find MITEs with weak TIR base-matching. In the present study, we developed a new strategy (TS clustering) for the systematic identification and classification of T2-MITEs on the basis of differences in short terminal sequences. This procedure is advantageous with respsct to MAK and FINDMITE as it can identify MITEs even in cases of weak TIR base-matching. By using this procedure, we identified 34398 T2-MITE candidates and classified them into 16 major subfamilies (TS subfamilies). The respective subfamilies were characterized on the basis of intra- and inter-cluster comparisons to population size, sequence diversity, and structural conservation. The evolutionary aspects of these subfamilies will be discussed.

Methods

Genome database and programming language

The *X. tropicalis* genome database (JGI version 4.1) was downloaded from the website of the Department of Energy (DOE) Joint Genome Institute (JGI) (genome.jgi-psf.org/Xentr4.home.html). All the scripts used for the present analyses were written in the Ruby language (version 1.8). The source codes used for searching and classifying the T2-MITEs will be provided upon request.

Calculation of pairwise alignment scores and pairwise sequence identities

Pairwise alignment scores and pairwise sequence identities (%) were calculated by using the stretcher program of the EMBOSS package (Rice et al. 2000). The distribution of the scores was figured using the statistical environment R (R Development Core Team 2008).

Extraction of a consensus sequence

A consensus sequence was obtained by calculating the frequencies of base usage at respective nucleotide positions of given sequences. The nucleotide base showing a frequency of 90% or more was defined as the consensus nucleotide at the respective nucleotide position. If the frequency was less than 90%, all the abundant nucleotide bases showing a total of more than 90% frequency were defined as the consensus nucleotide (for an example, if the frequency at a nucleotide position was 70% for A, 15% for C, 10% for G, and 5% for T, the consensus nucleotide base was V (A, C, or G).

Estimating genome-wide conservation of T2-MITE structure

In the present study, the extreme terminal 21-nucleotide (nt) sequence (TS stretch) and TSD sequence

(TTAA) of T2-MITE were investigated to determine their degree of conservation in the genome. For this aim, we used consensus sequences of TS stretches extracted from the respective T2-MITE clusters. The genome database was searched for each consensus TS stretch by regular expression (for example, AGGAACAGTAACAC[CT]AAAAAA for a TS stretch of the MITE cluster A1; see Table 3). The TSD sequence (TTAA) was omitted from this search. The obtained TS stretch sequences were divided into 2 groups depending on whether their sequences were present in the MITE structure (i.e., inverted pair of 2 TS stretches located within 1000 bp) or not, and the conservation extent was represented as a ratio of the number of TS stretches in the MITE structure to the total number of TS stretches in the genome. Furthermore, the 4-nt sequences externally connected to the TS stretches at both ends of the MITE structure were collected and used for TSD conservation analysis. The extent of TSD conservation was represented as a ratio of the number of intact TSD sequences to the total number of intact and mutated TSD sequences.

Selecting typical MITE sequences from the respective MITE clusters

To extract a typical MITE sequence representing each MITE cluster, we carried out a tournament selection of the MITE sequences by pairwise alignment analysis using the stretcher program of the EMBOSS package (Rice et al. 2000). Namely, the member sequences were randomly divided into small groups, which consisted of 10 or 11 sequences, and the highest total pairwise alignment score sequence in each group was selected and used for the next round of alignments. For example, 7326 sequences in MITE cluster A were divided into 726 groups of 10 sequences and 6 groups of 11 sequences. By pairwise alignment analysis, the 732 sequences with the highest total alignment scores were selected from the respective groups. These sequences were randomly divided into 71 groups of 10 sequences and 2

groups of 11 sequences, and then the 73 sequences with highest total alignment scores were reselected. By repeating this process for the above 73 sequences, we selected a typical MITE sequence from each MITE cluster. Only positive pairwise alignment scores were used to calculate the total alignment scores.

Dotplot analysis

Sequence similarity was analyzed by dotplot analysis using the polydot program of the EMBOSS package (Rice et al. 2000) with a word size of 8. For intra-cluster comparisons, the direction of each MITE sequence was sorted before the analysis, that is, the direction (forward or reverse) that yielded the highest alignment score to the representative sequence was chosen.

Searching databases of *Xenopus* repetitive sequences for MITEs

The Genbank database was searched for *Xenopus laevis* sequences homologous to *X. tropicalis* MITEs by the BLASTN program (on April 20, 2009: NCBI BLAST web site, www.ncbi.nih.gov/blast/) using the representative sequences as queries. The vrtrep.ref data of RepBase 14.03 (Jurka et al., 2005: www.girnst.org) was also searched to find known MITEs.

Results

Searching the genome database for T2-MITE candidates

The *X. tropicalis* genome database was searched for a combined sequence (T2 motif) of the target site (TTAA) and the terminal motif (AGGRR) present in the T2-MITE family (step 1 of Fig.1), by the regular expression search (TTAAAGG[AG][AG]). We then extracted the MITE candidates (sequences framed by an inverted pair of 2 T2-motifs) whose lengths ranged from 51 to 1000 bp (i.e. the underlined sequence of TTAAAGGRR...[41–990 bp]...YYCCTTTAA) because MITEs are generally shorter than 1000 bp. To eliminate "reduplicated" sequences, which were assumed not to reflect unique transpositions, but rather artificial duplications in the database, chromosomal DNA duplications, or replicative transpositions riding on another transposable element, the externally flanking 10-bp sequences of all candidate sequences were mutually compared. Therefore, all the candidate MITEs identified in the present study should be located at different chromosomal sites. In addition, sequences containing the indefinite nucleotide "N" were excluded from the analysis.

The T2 motif and its complementary sequence would be stochastically found at approximately 83,000 sites in a random nucleotide sequence of 1.7 billion bp (40% GC content), equivalent to the available *X. tropicalis* genome data (JGI version 4.1). We found 143,525 T2 motifs and 143,677 complementary T2 motifs (total, 287,202), indicating that T2 motifs were enriched 3.5 fold in the *X. tropicalis* genome. Among these motifs, 68,796 were organized into the T2-MITE configuration, meaning that there were 34,398 T2-MITE candidates in the available database. These T2-MITE candidates were used for the following classification.

Clustering of terminal sequences of T2-MITE candidates

Since the terminal sequence of a MITE is considered to interact with the transposase responsible for its transposition (Dufresne et al. 2007; Miskey et al. 2007; Yang et al. 2007), it is likely that all members within a certain MITE subfamily share an identical or closely-related inverted pair of terminal sequences. Hence, we classified the candidate MITEs on the basis of the differences in their terminal sequences. Since the TSD was the same for all the candidate MITEs, a short terminal sequence (TS stretch) that stretched from the TSD was used for clustering (step 2 of Fig. 1). We collected pairs of TS stretches from the respective MITE candidates and then grouped them into possible clusters according to the newly developed TS-clustering method as follows (step 3 of Fig. 1). If two TS stretch pairs were mutually identical or differed by only one base substitution, they were defined as belonging to a single cluster. In addition, if another TS stretch pair was identical to or differed by one base from at least one member of the cluster, we considered that this TS stretch pair also belonged to the same cluster. This is a reasonable assumption for those sequences derived from a common ancestor terminal sequence, because each of the amplified terminal sequences is independently mutated by random base substitution over a long evolutionary period. Actually, an arbitrary sequence present in a pool of TS stretch-pair sequences was used as a seed query to collect all members of a cluster. The search was carried out by allowing a one base-substitution of the query. The hit sequences in the pool were deposited into a tentative cluster (the first round collection). The members of the first tentative cluster were then used as queries to search the rest of the sequences from the pool and the hit sequences were deposited into another tentative cluster (the second round collection). The members of the second tentative cluster were used as queries for the next round collection and, in this way, the collection was recursively repeated. When no new members could be collected, the seed query and all the hit sequences were deposited as members of a final cluster. The unclassified sequences that remained in the pool were further used for clustering in the

same way as described above, and the clustering was repeated until there were no unclassified sequences present in the pool. The TS-clustering method was able to sort the MITE candidates into distinct clusters. In the present study, we focused on the major clusters consisting of more than 100 members.

It was likely that the extreme terminal sequence recognized by a transposase was common among members of a certain MITE subfamily, but it was unknown what length of TS stretch was necessary for TS clustering. We carried out TS clustering on various lengths of TS stretches to identify the minimum length of TS stretch that would lead to a reasonable result (Table 1). Individual TS clusterings (on different lengths of TS stretch pairs) were carried out on all the MITE candidates as a subject population. When a TS stretch length of 16 nucleotides (nt) (i.e., a TS stretch pair length of 32 nt) was used for TS clustering, the TS stretch pairs were grouped into 13 major clusters: *a* to *m*. Although members of clusters *l* and *m* decreased to below 100 (minor clusters) upon clustering with TS stretch lengths of 18 and 19 nt, respectively, all the other clusters were stable. When a length of 15 nt was used for TS clustering, the clusters *d*, *h*, and *l* were not separated. However, as described later, these clusters had distinguishable sequence identities. Thus, we classified the MITE candidates into 13 major clusters (A to M) according to the results obtained upon TS clustering with a 16-nt length; TS clusters of *a* to *m* corresponded to the MITE clusters A to M. Further analysis on MITE cluster M revealed that its member sequences contained some truncated segments homologous to members from MITE cluster A (data not shown). They did not show an inverted structure as a whole but were assembled from cluster A-derived sequences. Therefore, we determined that cluster M members were not MITEs. Thus we found 12 major MITE clusters of A to L corresponding to the 12 major TS clusters of *a* to *l*, respectively. Finally, 62% (21281/34398) of the MITE candidates were classified into the 12 major MITE clusters (Table 1).

Overall sequence similarity within each MITE cluster and among MITE clusters

To evaluate the present classification system, we investigated the intra- and inter-cluster similarities of the MITE sequences. Pairwise-alignment scores were calculated for 101 sets of 2 sequences selected randomly from each MITE cluster (intra-cluster alignment) or from 2 different MITE clusters (inter-cluster alignment). All MITE clusters showed higher alignment scores in the intra-cluster alignment than in any inter-cluster alignments, suggesting the validity of the present clustering procedure (Fig. 2A). In most intra-cluster pairwise alignments, the majority of sequence pairs yielded positive alignment scores, suggesting that each cluster members had a significant mutual similarity. However, these intra-cluster alignment scores were relatively low in cluster B in comparison with the other clusters: the median score for cluster B was -87 whereas those for the other clusters were from 248 to 2011. The lower score was considered to reflect the high sequence diversity in cluster B. Therefore, we characterized all the MITE clusters by their total length and GC content. Fig. 3A shows two-dimensional scatter plots of total length vs. GC content for 100 members randomly selected from each cluster. Cluster B showed apparently higher variation in GC content than the other MITE clusters, and its plot distribution had at least two focal points. In contrast, all the plot distributions for the other MITE clusters had only one focal point. These results suggested that members of cluster B would be heterogeneous and consist of two or more subpopulations.

Reclustering of the members of MITE clusters A and B

When TS clustering was carried out on 21-nt TS stretches, the members of TS cluster *b* were subdivided into four major TS clusters of *b*1 to *b*4, and the members of TS cluster *a* were subdivided into

two major clusters of *a*1 and *a*2 (Table 1). However, TS clustering on 22-nt TS stretches resulted in the omission of TS cluster *b*4 because many members (45/104) had fallen into minor clusters. Although the TS clusters *b*1 and *b*3 were further divided by TS clustering on the basis of TS stretch lengths of 25 and 23 nt, respectively, their MITE members could not be divided by differences in overall sequence (data not shown). Therefore, we regrouped MITE cluster A into A1 and A2 and MITE cluster B into B1 to B4 according to the results of TS clustering on the basis of TS stretch lengths of 21 nt. To evaluate the validity of this approach, we calculated the intra- and inter-cluster alignment scores for the newly-defined MITE clusters of A1, A2, B1, B2, B3, and B4, and found that they showed higher alignment scores in the intra-cluster alignment than those of the inter-cluster alignment (Fig.2B). We also made scatter plots of GC content vs. total length for 100 members from each MITE cluster and found that all the clusters showed one focal point in the plot distribution (Fig. 3B). Clusters B1 and B2 corresponded to the low-GC focal point, and cluster B3 corresponded to the high-GC focal points seen in Fig. 3A. High diversity in the length of the members from cluster B4 was due to tandem elongation of short internal sequences as described later.

Selection and characterization of the representative sequence in each MITE cluster

We examined the MITEs to select a representative sequence that could integrate all the members of each MITE cluster in sequence similarity. This could be accomplished if a sequence with the highest total alignment score could be selected from the respective MITE clusters. However, because there were too many MITE cluster members to calculate all the scores by one-on-one alignments, the highest score sequence was determined by the tournament selection method (see Methods; the representative sequences are available as Supplementary Material). Such a sequence was thought to be closely similar to its

ancestral MITE at the time of origin or amplification burst. The representative sequences obtained were characterized by the dotplot method (Fig. 4A; high resolution figures are available as Supplementary Material). Most of the representative sequences were considerably different from each other. The representative sequence of cluster A was identical to that of cluster A1, and the representative sequence of cluster B was identical to that of cluster B1. Interestingly, the left (~170 nt) and the right arms (~100 nt) of the B1 representative sequence were similar to the left arm of the B4 representative sequence and the right arm of the B2 representative sequence, respectively. In addition, the B4 representative sequence had a region consisting of tandem repeats of an internal sequence.

The dotplot analysis also revealed that the representative sequences of clusters A1, A2, B3, and J were inverted-repeat structures over their whole length (nearly perfect inverted-repeat), while inverted-repeat structures in other clusters were restricted to the short terminal stretches. Sequence similarities between the left and right TS stretches in each MITE cluster were also characterized by the base-matching ratio between the left and right TS stretches (the ratio for perfect base-matching is 1.0; Table 2). High base-matching ratios (0.88 or more) were observed in clusters A1, A2, B3, and J, while low base-matching ratios (0.6 or less) were observed in clusters I, F, K, and B2. In addition, cluster K showed the lowest ratio (0.75) even with a TS stretch length of 11 nt, indicating that its left and right terminal sequences were mutually different.

Sequence "uniformity" in each MITE cluster

We investigated the uniformity among the members of each MITE cluster. As an example, 9 sequences randomly selected from each MITE cluster were compared with the representative sequence of the corresponding MTIE cluster by the dotplot method (Fig. 4B). We observed that "uniformity" was

considerably varied among the MITE clusters. Clusters A1 and C showed the highest levels of uniformity judging from their dotplot patterns, although they included diverged members at low frequencies. In addition, completely identical copies were found in clusters A1, B3, and C. Cluster A1 included 6 sets of 3 identical copies and 79 sets of 2 identical copies. Cluster B3 included 1 set of 4 identical copies, 3 sets of 3 identical copies, and 1 set of 2 identical copies. Cluster C included 1 set of 6 identical copies, 1 set of 3 identical copies, and 3 sets of 2 identical copies. In contrast, members from cluster E showed the lowest uniformity. It should also be noted that some members of cluster B4 included tandemly elongated internal sequences and contributed to the high variation in MITE length as seen in Fig. 3B.

In addition, we investigated the pairwise sequence identity (%) between each member and its representative sequence (Fig. 5). The members of clusters A1 and C showed the highest levels of sequence identity to the corresponding representative sequence; 88% of the members showed more than 80% sequence identity to the representative sequence. In contrast, the members of clusters E and H showed relatively low levels of sequence identity to the corresponding representative sequence. Even in these clusters, however, more than 65% of members showed more than 60% sequence identity to the representative sequence. The members of cluster B4 showed lower identities than those of the other clusters, which could be a result of the high variation in the MITE length, as descrived above.

Genome-wide aspect of the conservation of MITE structure and TSD

Severely mutated MITEs must be present in the genome, and some of them were considered to have lost their inverted-repeat structure and/or TSD sequence. Thus, such mutated MITEs would not be harvested in the present collection. Therefore, we investigated the number of TS stretches specific to

each MITE cluster present in the genome and the fraction of TS stretches that were organized as a part of the MITE structure. This fraction would reflect the extent of MITE structural conservation during evolution. Thus, we searched the genome database for consensus sequences of TS stretches for the respective MITE clusters (Table 3) and identified TS stretches that were organized into the MITE structure (see Methods). For this analysis, 21-nt-long TS stretches were used, because some shorter TS stretches from clusters C and K were indistinguishable (for example, the 20-nt sequence of AGGAGACATATTATATAAAT matched both C and K). In addition, clusters B1, B2, and B4 were analyzed as the combined cluster B1/2/4 by a consensus TS stretch sequence, because they shared the left and/or right TS stretch. The results were shown as a ratio of the number of TS stretches organized into the MITE structure to the total number of TS stretches found in the database (Table 3, Fig. 6A). The highest ratios (0.48–0.50) were obtained for clusters A1 and C, whereas the lowest ratios (0.07–0.12) were obtained for clusters G, H, E, F, and I, indicating that the TS stretches of clusters A1 and C were well conserved as MITE structures in the genome, but those for clusters G, H, E, F, and I were not. It could be suggested that the members of clusters G, H, E, F, and I underwent severe mutation and were eventually destroyed during evolution.

The TSD sequence was considered to be TTAA at the time of origin, but mutations accumulated over time. Therefore, we calculated the ratio of the number of intact TSDs to the total number of intact and mutated TSDs by collecting all the 4-bp sequences connected to the respective TS stretches that were organized into the MITE structure (Table 3, Fig. 6B). The highest ratios (0.92–0.98) were obtained for clusters A1, A2, B3, C, and K. In contrast, cluster E showed significantly low ratios (0.60). Figure 6C shows the correlation between the conservation ratios of MITE structure and TSD (correlation coefficient, 0.74). In particular, both ratios were extremely high in clusters A1 and C, suggesting that they might

include active or at least recently active MITEs for transposition and/or amplification (see Discussion).

Searching the MITE clusters for known MITEs

By using BLASTN to search the database (see Methods), we found that 10 of the 16 MITE clusters had high sequence similarities with known *X. laevis* and *X. tropicalis* transposons (Table 4). Clusters A1, B2, E, F, H, and J corresponded to Xmix (Hikosaka and Kawahara 2004), PIR, XBR, REM1, XR, and XFB (Ünsal and Morgan 1995), respectively. Clusters B1 (B4), C, and L were homologous to T2_1, T2_2a, and T2_3 of *X. tropicalis* MITEs (RepBase: Jurka et al. 2005), respectively. We could not find any homologs for clusters A2, B3, D, G, I, and K in the database. These results showed that the present TS-clustering method worked well in the classification of MITEs.

Conservation of the MITE clusters in lineages of *Xenopus* species

Our investigations indicated that homologs of the 12 clusters were found in *X. laevis*, while there were no counterparts for clusters A2, B1, B4, and K (Table 4). Although clusters A2 and K showed high homologies to 2 *X. laevis* sequences that were present in the 3′-untranslated region (UTR) of the mRNAs for MGC83628 and MGC84184, respectively; these were not repetitive sequences but unique sequences in the *X. laevis* genome. Although some *X. laevis* sequences showed significant homology to the approximately 100-nt stretch of the right arm of cluster B1, all these sequences exhibited significantly greater homology to cluster B2. Since the right arm sequence of B2 was shared by the right arm of B1 (see Fig. 4A), these "hit" sequences were considered to be *X. laevis* counterparts of cluster B2 and not cluster B1. Cluster B4 showed no significant homology to any *X. laevis* sequences.

Discussion

MITEs are a universal component of living organisms and must have contributed to their evolution. Genome-wide searches for MITEs and their classification are necessary to obtain insight into the evolutionary dynamics of the creation and turnover of MITEs. In the present study, we carried out an exhaustive search, classification and characterization of *X. tropicalis* T2-MITEs. To search the genome database for T2-MITEs, we used only a short terminal sequence (T2 motif) characterizing T2-MITEs and a defined length of less than 1000 bp. Therefore, we did not know the internal sequences of any of the candidate T2-MITEs identified. Nevertheless, most of the candidates appeared to be T2-MITEs. There are two types of procedures to identify MITE from genome databases, the computer programs MAK (Yang and Hall 2003) and FINDMITE (Tu 2001). The former depends on a homology search for known MITEs. Therefore, it would not be able to find new MITEs that have weak or no similarity to the reference MITE. Although the latter does not depend on sequence information of known MITEs and utilizes TSD sequences and TIR base matching thresholds to find MITEs, the utility of the TIR base matching would restrict its capacity to find MITEs that have weak TIR base matching. On the other hand, the present method (TS clustering) is capable of finding MITEs even in cases of weak TIR base matching (e.g., cluster K). Therefore, our present T2-MITE collection consists of a variety of T2-family MITE members.

The use of a short stretch connected to TSD enabled us to classify the candidates into well-defined MITE clusters, some of which actually corresponded to known MITE subfamilies. We found 12 major MITE clusters (A–L) on the basis of differences in 16-nt-long TS stretches. Furthermore, we found 2 clusters (A1 and A2) from cluster A and 4 clusters (B1, B2, B3, and B4) from cluster B on the basis of differences in 21-nt-long TS stretches. Major members of each MITE cluster had strong similarities in

their internal sequences to the representative sequence, but some minor members only showed a weak identity (see Fig. 5). It is considered that all these members could be created by a common transposition/amplification machinery, because MITEs with similar terminal sequences, despite differences in their internal sequences, could be mobilized by a single transposase or closely-related transposases (Zhang et al. 2001). In general, families and subfamilies of transposable elements have been defined by their overall sequence similarity (Wicker et al. 2007), while their classification system is still an open problem (Kapitonov and Jurka 2008; Seberg and Peterson 2009). Since we used only the TS stretch for the classification of MITEs, the MITE clusters could not be equated with a "subfamily" defined by overall sequence similarity. Therefore, we propose that the present MITE clusters are called "TS subfamilies," each of which includes at least one MITE subfamily defined by overall similarity to the representative sequence. In the present study, we were able to define 16 major TS subfamilies and their corresponding MITE subfamilies (10 known and 6 novel subfamilies) in the *X. tropicalis* T2-MITE family. Thus, the results reveal that a short TS stretch sequence is also useful to classify MITEs into well-defined subfamilies. It might mean that the extreme terminal sequence of MITEs is indispensable to form the respective subfamilies, and each subfamily might be created by a specific transposase that appeared during evolution.

Considerations on the evolutionary aspects of the 16 T2 MITE subfamilies

Repetitive sequences homologous to 12 of the 16 representative sequences of *X. tropicalis* TS subfamilies were found in *X. laevis* (see Table 4), suggesting that these subfamilies originated before the branching of the 2 species, because it is unlikely that 1 of the 2 *Xenopus* species acquired multiple MITE subfamilies by "horizontal transfer" and that their transfers occurred simultaneously with transfers of the

transposase genes. In addition, our previous study showed that copies of Xmix (subfamily A1) were clearly discrete between the two *Xenopus* lineages with many lineage-specific modifications (Hikosaka and Kawahara 2004), suggesting their vertical inheritance. The lineage of *X. tropicalis* had branched from the other *Xenopus* species, approximately 60 (Evans et al. 2004) or 100 million years ago (Knochel et al. 1986). Therefore, it is likely that at least some subfamilies have been inherited over along period of time (more than 60 million years) through vertical inheritance. In contrast, cluster B4 showed no significant homology to any *X. laevis* sequences, suggesting that it originated in the *X. tropicalis* lineage after its branching from the *X. laevis* lineage.

Large fractions (~50%) of TS stretches from TS subfamilies A1 and C were organized as MITE structures, whereas a very small fraction (7–12%) of TS stretches from the TS subfamilies E, F, G, H, and I were organized as MITE structures (see Table 3). The intact TSD exists at high conservation ratios in the TS subfamilies A1, A2, B3, C, and K than for the other TS subfamilies (see Table 3). These ratios are as high as those of the *Xenopus* Uribo1 and Uribo2 transposons (data not shown), which are thought to be recently active (Hikosaka et al 2007). MITE sequences are highly uniform in the TS subfamilies A1 and C but not in E, G, and H (see Fig. 4B). In particular, TS subfamilies A1, B3 and C included completely identical copies. In general, an amplification burst of MITEs is thought to occur one or more times after its origin; therefore, the most recent burst would increase the uniformity of MITE members. Taking these into consideration, we postulate that the TS subfamilies A1, B3, and C might be active or at least recently active in transposition and amplification, whereas the TS subfamily E might have lost its activities a long time ago.

There are many questions to be elucidated concerning the creation, propagation, and turnover of MITEs in the genome. Particularly, the extensive longevity of the TS subfamilies A1 and C in a host is

interesting because, in general, DNA transposons are thought to be ephemeral in a host species (for reviews, see Brookfield 2005; Miskey et al. 2005). These TS subfamilies will be optimal subjects to study the molecular mechanisms and evolutionary dynamics of T2-MITE amplification. The information obtained in the present study provides valuable clues for future studies on the role of T2-MITEs in the evolution of genomes.

Acknowledgements

We thank the DOE JGI for providing the *X. tropicalis* genome data. This work was supported by a Grant-In-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology.

References

Avramova Z, Tikhonov A, Chen M, Bennetzen JL (1998) Matrix attachment regions and structural colinearity in the genomes of two grass species. Nucleic Acids Res 26:761-767

Brookfield J (2005) The ecology of the genome — mobile DNA elements and their hosts. Nat Rev Genet 6:128-136

Brügger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA (2002) Mobile elements in archaeal genomes. FEMS Microbiol Lett 206:131-141

Bureau TE, Wessler SR (1992) *Tourist*: a large family of small inverted repeat elements frequently

associated with maize genes. Plant Cell 4:1283-1294

Dufresne M, Hua-Van A, El Wahab HA, M'Barek SB, Vasnier C, Teysset L, Kema GHJ, Daboussi MJ (2007) Transposition of a fungal miniature inverted-repeat transposable element through the action of a *Tc1*-like transposase. Genetics 175:441-452

El Amrani A, Marie L, Aïnouche A, Nicolas J, Couée I (2002) Genome-wide distribution and potential regulatory functions of *AtATE*, a novel family of miniature inverted-repeat transposable elements in *Arabidopsis thaliana*. Mol Genet Genomics 267:459-471

Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC (2004) A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. Mol Phylogenet Evol 33:197-213

Feschotte C, Osterlund MT, Peeler R, Wessler SR (2005) DNA-binding specificity of rice *mariner*-like transposases and interactions with *Stowaway* MITEs. Nucleic Acids Res 33:2153-2165

Feschotte C, Zhang X, Wessler RW (2002) Miniature inverted-repeat transposable elements and their relationships to established DNA transposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, D.C., pp 1093-1110

Hikosaka A, Kawahara A (2004) Lineage-specific tandem repeats riding on a transposable element of

MITE in *Xenopus* evolution: a new mechanism for creating simple sequence repeats. J Mol Evol 59:738-746

Hikosaka A, Kobayashi T, Saito Y, Kawahara A (2007) Evolution of the *Xenopus piggyBac* transposon family TxpB: domesticated and untamed strategies of transposon subfamilies. Mol Biol Evol 24:2648-2656

Hikosaka A, Yokouchi E, Kawahara A (2000) Extensive amplification and transposition of a novel repetitive element, Xstir, together with its terminal inverted repeat in the evolution of *Xenopus*. J Mol Evol 51:554-564

Izsvák Z, Ivics Z, Shimoda N, Mohn D, Okamoto H, Hackett PB (1999) Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J Mol Evol 48:13-21

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O. Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogent Genome Res 110:462-467

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements inplemented in Repbase. Nat Rev Genet 9:411-412

Knöchel W, Korge E, Basner A, Meyerhof W (1986) Globin evolution in the genus *Xenopus*:

comparative analysis of cDNAs coding for adult globin polypeptides of *Xenopus borealis* and *Xenopus tropicalis*. J Mol Evol 23:211-223

Loot C, Santiago N, Sanz A, Casacuberta JM (2006) The proteins encoded by the *pogo*-like *Lemi*1 element bind the TIRs and subterminal repeated motifs of the *Arabidopsis Emigrant* MITE: consequences for the transposition mechanism of MITEs. Nucleic Acids Res 34:5238-5246

Miskey C, Izsvák Z, Kawakami K, Ivics Z (2005) DNA transposons in vertebrate functional genomics. Cell Mol Life Sci 62:629-641

Miskey C, Papp B, Mátés L, Sinzelle L, keller H, Izsvák Z, Ivics Z (2007) The ancient *mariner* sails again: transposition of the human *Hsmar1* element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. Mol Cell Biol 27:4589-4600

Naito K, Cho E, Yang G, Campbell M, Yano K, Okumoto U, Tanisaka T, Wessler SR (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103:17620-17625

Piriyapongsa J, Jordan IK (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE 2:e203

Piriyapongsa J, Mariño-Ramírez L, Jordan IK (2007) Origin and evolution of human microRNAs from

transposable elements. Genetics 176:1323-1337

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 Available via DIALOG. http://www.R-project.org.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277

Seberg O, Petersen G (2009) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet 10:276

Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc Natl Acad Sci USA 98:1699-704

Ugarkovic D (2005) Functional elements residing within satellite DNAs. EMBO Rep 6:1035-1039

Ünsal K, Morgan GT (1995) A novel group of families of short interspersed repetitive elements (SINEs) in *Xenopus*: evidence of a specific target site for DNA-mediated transposition of inverted-repeat SINEs. J Mol Biol 248:812-823

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic

transposable elements. Nat Rev Genet 8:973-982

Yang G, Hall TC (2003) MAK, a computational tool kit for automated MITE analysis. Nucleic Acids Res 31:3659-65

Yang G, Zhang F, Hancock CN, Wessler SR (2007) Transposition of the rice miniature inverted repeat transposable element *mPing* in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 104:10962-10967

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR (2001) P instability factor: An active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposons. Proc Natl Acad Sci USA 98:12572-12577

Zhou F, Tran T, Xu Y (2008) *Nezha*, a novel active miniature inverted-repeat transposable element in cyanobacteria. Biochem Biophys Res Commun 365:790-794

Figure Legends

Figure 1 Schematic illustration of the extraction and classification of T2-MITEs

(1) The *X. tropicalis* genome database was searched for T2 motifs and then for candidate T2-MITEs that were framed by an inverted pair of two T2 motifs. (2) Pairs of short terminal sequences (TS stretch pairs) of the MITE candidates were collected for the following classification. TS stretches are shown by arrows. (3) TS stretch pairs were classified into TS clusters by the TS clustering method (see text). (4) MITE candidates were classified into MITE clusters according to the results obtained from TS clustering. Arrows and arrowheads indicate the 5′ to 3′ direction of the respective sequences.

Figure 2 Intra- and inter-MITE cluster similarities and differences revealed by pairwise alignment analysis

(A): One hundred and one sets of two MITE candidate sequences (randomly-paired two sequences) from each cluster (intra-cluster alignment) or from two different clusters (inter-cluster-alignment) were analyzed by pairwise alignment. Each graph shows the distribution of the scores for intra-cluster alignment (marked by a cluster name at the upper left side) and for inter-cluster alignments (marked at the X-axis by cluster names, one of which is the name used for the intra-cluster alignment). (B) The intra- and inter-cluster alignments were performed for clusters A1, A2, B1, B2, B3, and B4 in the same way as described above. Numerals inserted in each graph indicate the median score for the intra-cluster alignments.

Figure 3 Intra- and inter-cluster comparison of the MITEs by analyses of GC content and total length

One hundred sequences from each MITE cluster based on TS clustering with (A) 16 nt or (B) 21 nt

were analyzed for GC content and total length. The results are shown as scatter plot graphs of GC content vs. total length. Two numerals noted along the X- and Y-axes indicate the standard deviations for the means of total length and GC content, respectively. Letters inserted in the graphs indicate the names of the MITE clusters.

Figure 4 Intra- and inter-cluster comparison of the MITEs by dotplot analysis

(A) The representative sequences extracted from the respective MITE clusters (see Methods) were compared by polydot analysis. The comparisons were carried out for their forward (F) and reverse-complement (R) sequences. (B) The representative sequence of each MITE cluster (the extreme upper left sequence) and nine sequences selected from corresponding MITE clusters were compared by polydot analysis. Letters inserted in the graphs indicate the names of the MITE clusters.

Figure 5 Sequence identities of MITE cluster members

All MITE cluster members were aligned with the representative sequence, and their sequence identity levels (%) were calculated. The fraction of members that have identity levels of 0–20%, >20–40%, >40–60%, >60–80%, and >80–100% are shown by open, gray, dotted, hatched, and closed bars, respectively. Letters inserted in the graph indicate the names of the MITE clusters.

Figure 6 Conservation rates of MITE structure and TSD sequences in the genome

(A) Numbers of TS stretches in the MITE (closed bar) and non-MITE structures (open bar) were determined by searching the genome database (see Text); these numbers for the respective MITE clusters are shown as histograms. Letters inserted in the graph indicate the names of the MITE clusters. (B)

Numbers of intact and mutated TSD sequences were determined for the cluster-specific TS stretches in the MITE structure. Ratios of the number of intact TSD sequences to the total number of intact and mutated TSD sequences were calculated and are shown as histograms. (C) The correlation between the conservation ratios of MITE structure and TSD sequence are shown as a scatter plot graph. Letters inserted in the graphs indicate the names of the MITE clusters. Bold marks indicate the clusters containing identical MITE copies.

length (nt) of TS stretch	-	10	11	12	13	14	15	**16**	17	18	19	20	**21**	22	23	24	25
# TS stretch pairs clustered as major TS cluster members	(34398)	29947	27726	24160	18802	7459	7386	**7326 : a**	7234	7113	7034	6568	**6487 : a1**	6384	6217	6099	6009
												384	**346 : a2**	338	307	275	241
						10057	7502	**7120 : b**	6826	6582	6288	5765	**2563 : b1**	2520	2438	2386	2045
																	255
													1636 : b2	1518	1410	1288	1217
													1271 : b3	1251	1102	1077	1059
															128	127	126
												111	**104 : b4**				
							2128	**1267 : d**	1142	995	792	738	635 (d1)	594	491	428	386
								449 : h	385	346	311	267	237 (h1)	199	169	151	129
								161 : l	100								
						556	453	**408 : i**	367	307	253	212	177 (i1)	138	120		
					648	585	534	**487 : g**	448	387	352	292	263 (g1)	235	215	191	172
					1119	1049	942	**867 : f**	777	735	683	645	575 (f1)	494	413	355	271
					1599	1388	1254	**1142 : e**	934	703	473	333	233 (e1)	161	126	113	
					363	347	321	**306 : j**	295	287	279	260	244 (j1)	224	183	168	157
				1964	1913	1586	1550	**1530 : c**	1455	1441	1429	1417	1399 (c1)	1369	1310	1277	1225
						224	219	**218 : k**	203	199	193	188	184 (k1)	176	168	165	157
			127	119	114	112	111	**108** : m	105	104							

Table 1: Result of hierarchical clustering of 34408 TS stretch pairs. The letters in columns 16 and 21 are the name of major TS clusters.

MITE cluster	base-match ratio		
	11 nt	18 nt	25 nt
A	0.98	0.95	0.92
(A1)	0.98	0.96	0.93
(A2)	0.98	0.96	0.91
B	0.94	0.82	0.71
(B1)	0.94	0.83	0.70
(B2)	0.96	0.75	0.59
(B3)	0.94	0.92	0.92
(B4)	0.99	0.83	0.70
C	0.97	0.81	0.68
D	0.94	0.89	0.72
E	0.85	0.82	0.75
F	0.84	0.72	0.56
G	0.95	0.82	0.73
H	0.95	0.85	0.71
I	0.90	0.64	0.52
J	0.91	0.92	0.88
K	0.75	0.65	0.58
L	0.84	0.80	0.66

Table 2: Base-match ratios of left and right TS stretches.

MITE cluster	TS stretch (21 nt) consensus sequence	TS stretch			TSD		
		total	MITE-form	ratio	total	intact	ratio
A1	AGGAACAGTAACACYAAAAAA	15262	7556	0.50	7124	6946	0.98
A2	AGGAAAAGTAACRCTAAMWWW	1438	448	0.31	408	377	0.92
B1/2/4	AGGAGAARGAAAGKYWWWWWV	16523	4842	0.29	4760	4163	0.87
B3	AGGRRAARGAAAGKCWAAGTC	5353	1328	0.25	1102	1033	0.94
C	AGGAGACATATYSKATAAAWR	3213	1546	0.48	1512	1446	0.96
D	AGGRRAACTAYACCCCHVRRH	5533	926	0.17	908	733	0.81
E	AGGGGWDGTTYACYTTYDDDH	8605	860	0.10	856	513	0.60
F	AGGRATWCTGTCATGRKWWWW	4525	516	0.11	510	380	0.75
G	AGGGGACCTGTCACCYWVAVA	2774	200	0.07	198	171	0.86
H	AGGAGAACTAAASCYTAVHDW	2635	240	0.09	238	206	0.87
I	AGGGGAACTYYRSCYWMHVWM	2769	344	0.12	340	248	0.73
J	AGGARYAGTTCAGTGTAAAAA	1351	214	0.16	210	165	0.79
K	AGGRGAYATWTWVYATMMWKT	796	266	0.33	266	251	0.94
L	AGGGGWWVTAAACCCWRYBRY	1283	210	0.16	204	162	0.79

Table 3: Genome-wide conservation of TS stretch in MITE-form and intact TSD sequence.

MITE cluster	BLASTN search to *X. laevis* sequences (NCBI)					homology to known TE
	query (representative sequence)	min. e-value	#hits < 1e-10	query coverage		
A1	scaffold_84:1289835-1290251	3.00E-97	99	full length	100%	Xmix
A2	scaffold_71:3025925-3026357	5.00E-11	1	left + right	70%	
B1	scaffold_259:1603354-1603822(cmp)	2.00E-29	68	right (104bp)	22%	T2_1_Xt
B2	scaffold_410:1022327-1022808	2.00E-60	83	full length	95%	PIRd_Xt
B3	scaffold_90:2435402-2435815	3.00E-21	14	full length	97%	
B4	scaffold_512:446958-447825	0.026	0	left (28 bp)	3%	T2_1_Xt
C	scaffold_104:1689234-1689746	1.00E-138	17	full length	100%	T2_2a_Xt
D	scaffold_21:4521085-4521586	1.00E-138	79	full length	96%	
E	scaffold_470:347035-347493	6.00E-112	383	full length	100%	XBR_Xt
F	scaffold_32:707552-708165	1.00E-138	155	full length	99%	REM1_XL
G	scaffold_211:182448-182966	2.00E-125	81	full length	100%	
H	scaffold_938:267398-268006	1.00E-164	105	full length	99%	XR-b_Xt
I	scaffold_582:446911-447479	2.00E-148	24	full length	100%	
J	scaffold_164:1450463-1450959	9.00E-97	80	full length	100%	XFB_Xt
K	scaffold_11:1621369-1621974	2.00E-17	1	inner frag.	15%	
L	scaffold_1551:32919-33517	2.00E-60	13	inner frag.	60%	T2_3_Xt

Table 4: Results of homology search to *X. laevis* sequences and known transposable elements (TEs).

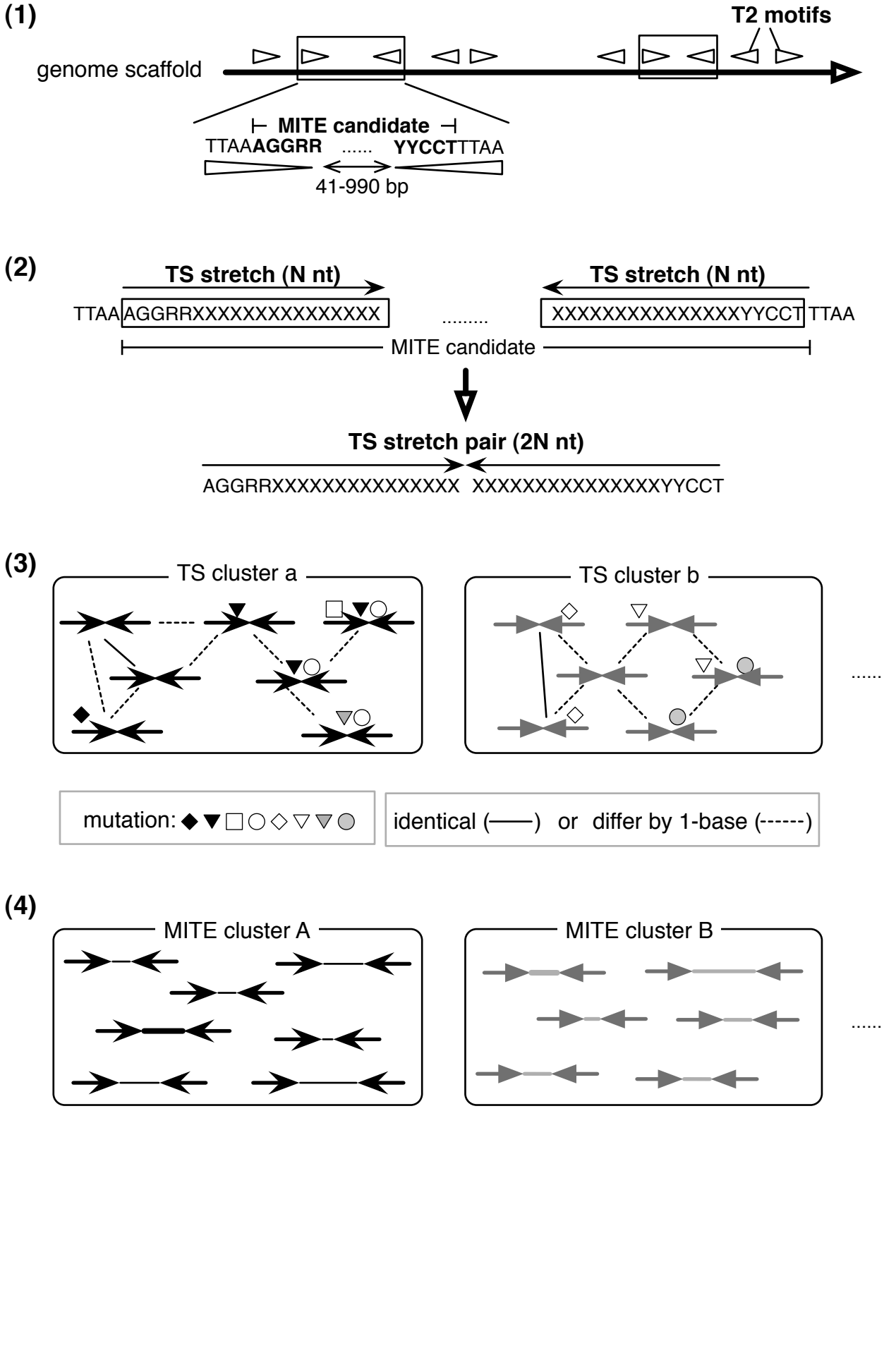
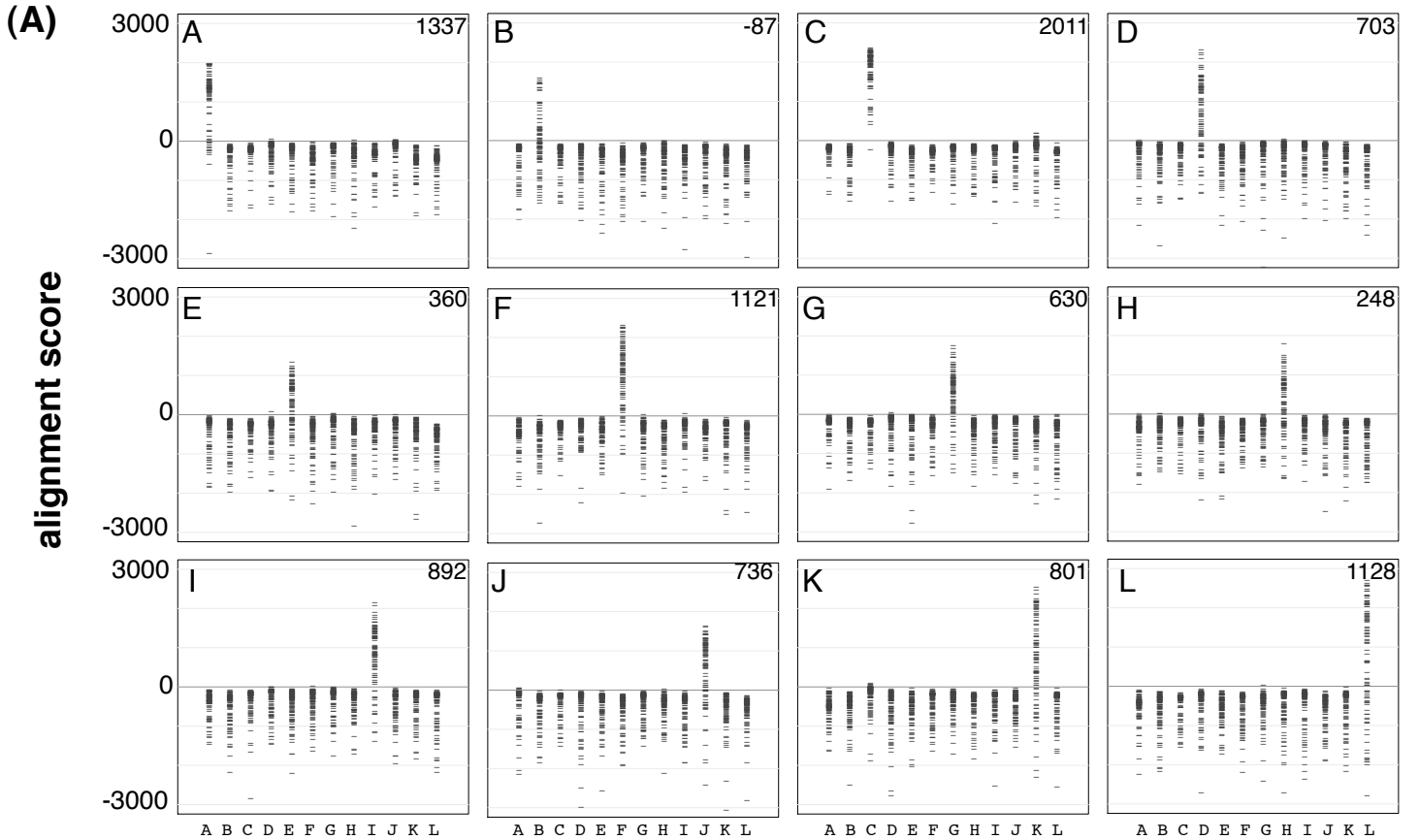

(1)
T2 motifs
genome scaffold
MITE candidate
TTAAAGGRR ..... YYCCTTTAA
41-990 bp
(2)
TS stretch (N nt)
TS stretch (N nt)
TTAA AGGRRXXXXXXXXXXXXXXX
.....
XXXXXXXXXXXXXXXYYCCT TTAA
MITE candidate
TS stretch pair (2N nt)
AGGRRXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXYYCCT
(3)
TS cluster a
TS cluster b
.....
mutation: ◆ ▼ □ ○ ◇ ▽ ▼ ●
identical (——) or differ by 1-base (------)
(4)
MITE cluster A
MITE cluster B
.....

Figure 1


(A)
alignment score
3000
0
-3000
A 1337
B -87
C 2011
D 703
E 360
F 1121
G 630
H 248
I 892
J 736
K 801
L 1128
A B C D E F G H I J K L

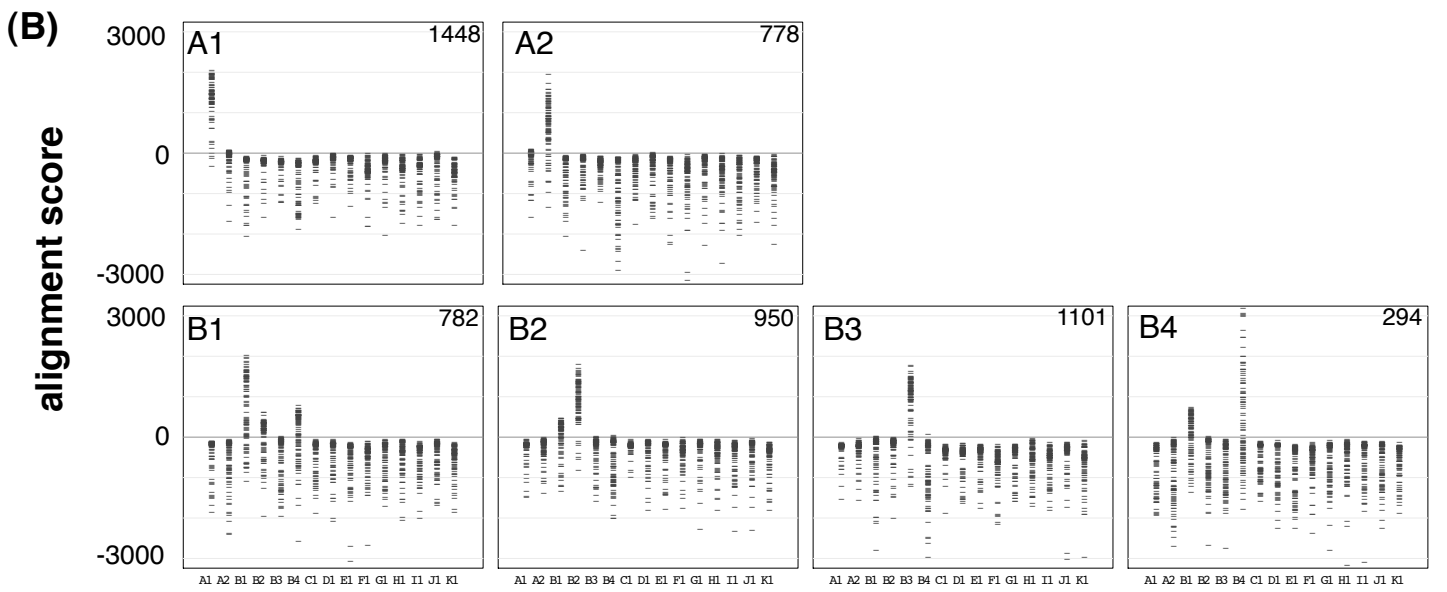

(B)
alignment score
3000
0
-3000
A1 1448
A2 778
B1 782
B2 950
B3 1101
B4 294
A1 A2 B1 B2 B3 B4 C1 D1 E1 F1 G1 H1 I1 J1 K1

Figure 2

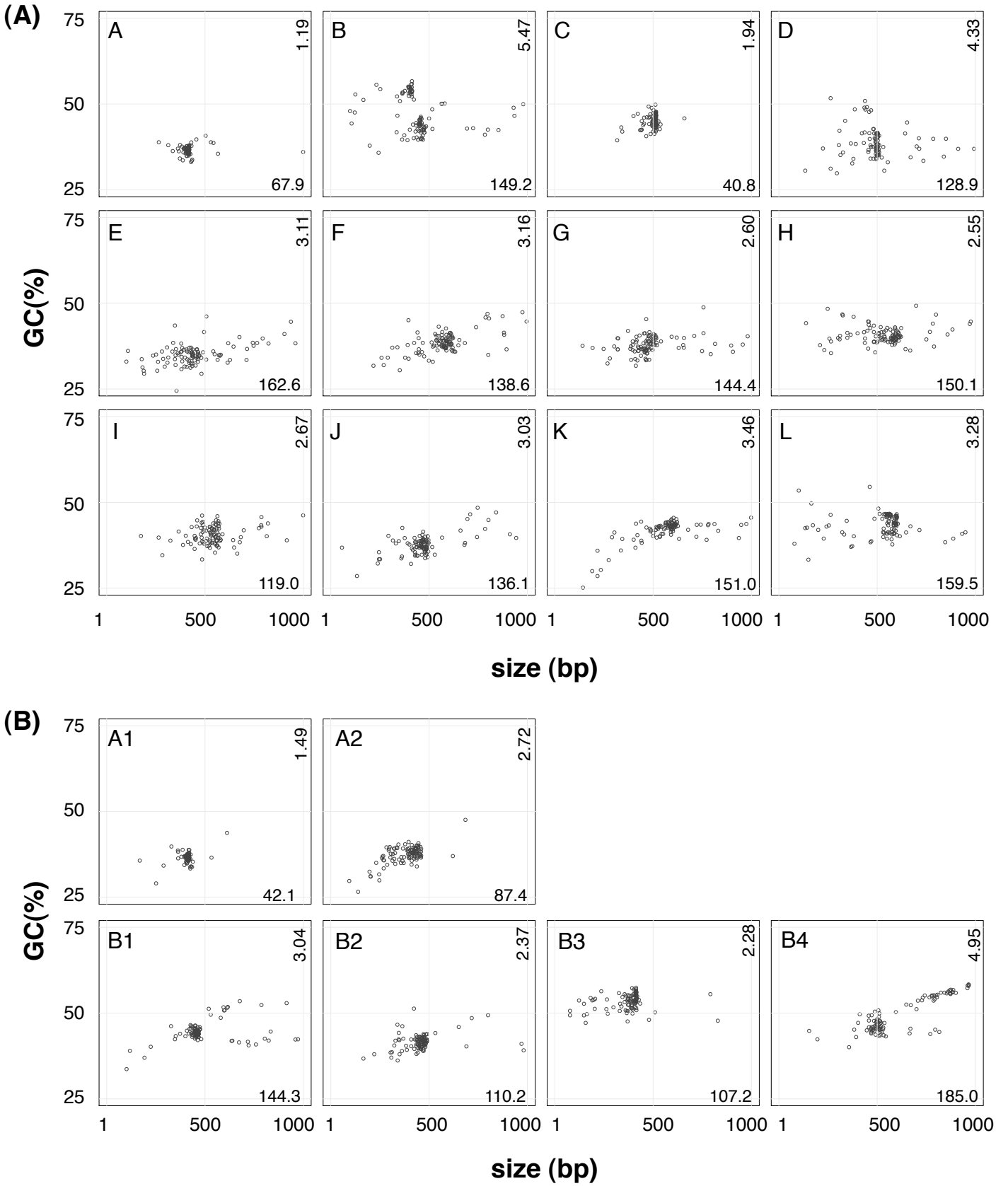

(A)
A
1.19
67.9
B
5.47
149.2
C
1.94
40.8
D
4.33
128.9
E
3.11
162.6
F
3.16
138.6
G
2.60
144.4
H
2.55
150.1
I
2.67
119.0
J
3.03
136.1
K
3.46
151.0
L
3.28
159.5
GC(%)
75
50
25
1
500
1000
size (bp)
(B)
A1
1.49
42.1
A2
2.72
87.4
B1
3.04
144.3
B2
2.37
110.2
B3
2.28
107.2
B4
4.95
185.0
GC(%)
size (bp)

Figure 3

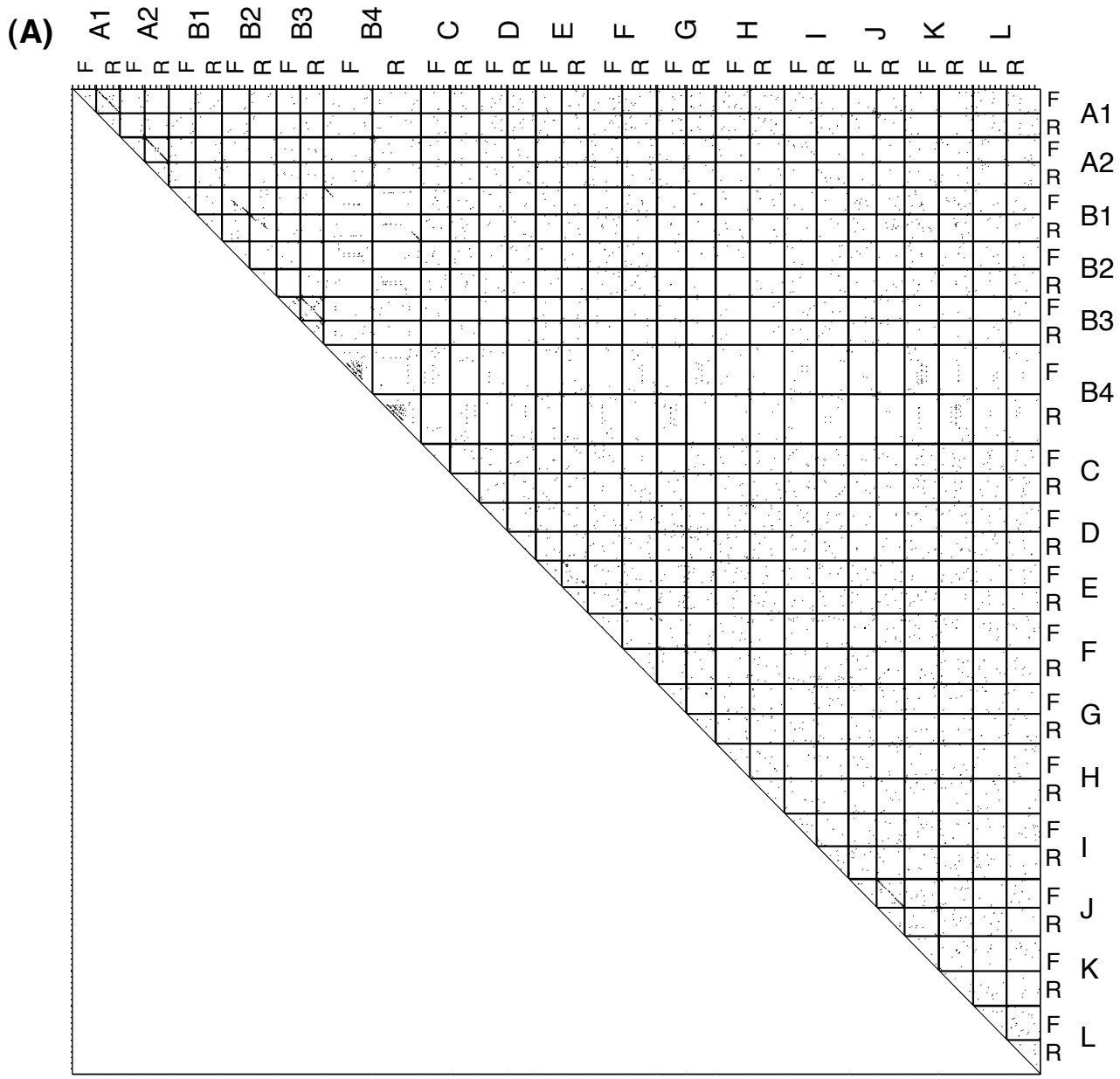


Figure 4

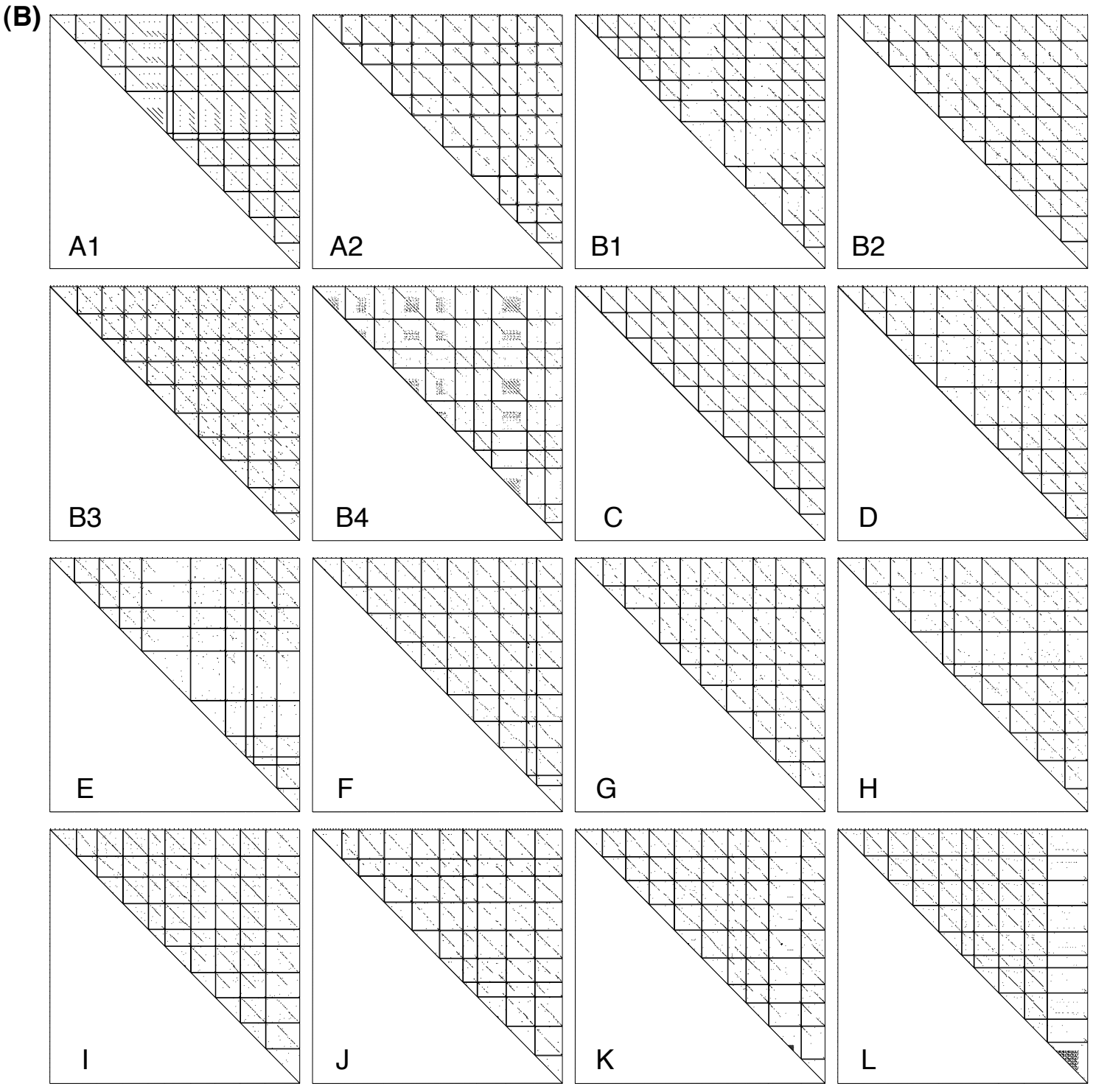


Figure 4

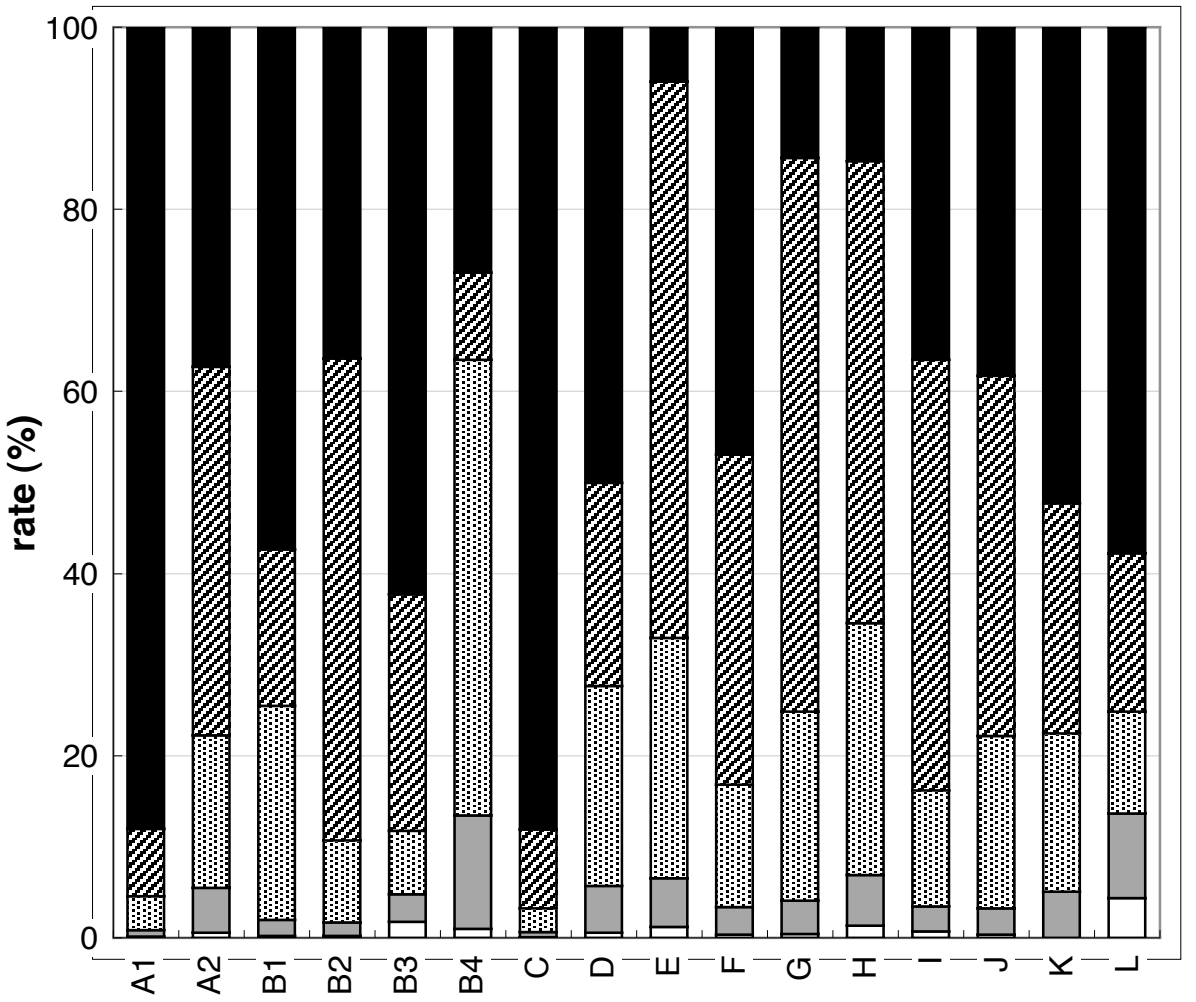


Figure 5

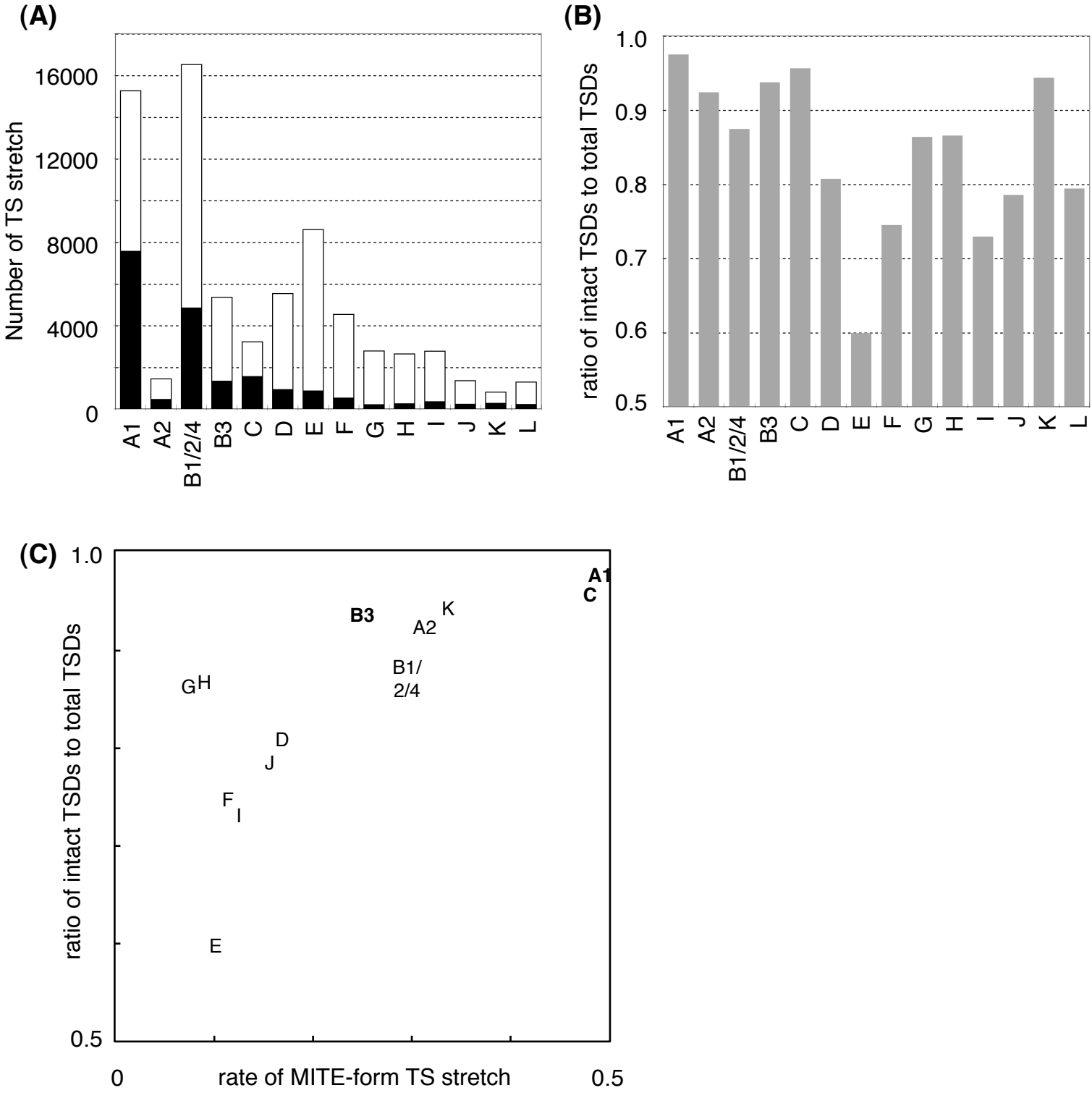

(A)
Number of TS stretch
16000
12000
8000
4000
0
A1
A2
B1/2/4
B3
C
D
E
F
G
H
I
J
K
L
(B)
ratio of intact TSDs to total TSDs
1.0
0.9
0.8
0.7
0.6
0.5
A1
A2
B1/2/4
B3
C
D
E
F
G
H
I
J
K
L
(C)
ratio of intact TSDs to total TSDs
1.0
0.5
0
rate of MITE-form TS stretch
0.5
A1
C
B3
K
A2
B1/
2/4
G
H
D
J
F
I
E

Figure 6