Evolutionary insights into Mariner-like elements in *Apis* species

Kakeru Yokoi[1,]*, Kiyoshi Kimura[2], Hidemasa Bono[3,4]

1 Insect Genome Research and Engineering Unit, Division of Applied Genetics, Institute of Agrobiological Sciences (NIAS), National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba, Ibaraki 305-8634, Japan

2 Honeybee Research Group, Division of Animal Breeding and Reproduction, National Institute of Livestock and Grassland Science (NILGS), National Agriculture and Food Research Organization (NARO), Tsukuba, 2 Ikenodai, Tsukuba, Ibaraki 305-0901, Japan

3 Database Center for Life Science (DBCLS), Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

4 Program of Biomedical Science, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima City, Hiroshima, 739-0046, Japan

*Corresponding author. E-mail: yokoi123@affrc.go.jp

Key words: *Apis mellifera*, *Apis cerana*, *Apis florea*, *Apis dorsata*, Mariner-like-elements, transposase.

Abstract

Mariner and mariner-like elements (MLEs) are distributed in various species and their sequences are highly diverse. In previous studies, transposable elements in *Apis* species mainly consisted of *mariner* and MLE. However, *Apis* MLE evolution has not yet been elucidated in detail. For it, MLEs were searched for in multiple *Apis* species, and those detected were then classified. More MLEs were detected in the *A. mellifera* genome than in other *Apis* species genomes, and this was not attributed to the domestication of *A. mellifera*. The MLEs detected were classified into 31 *Drosophila* MLE classes. In this classification, almost all MLEs were classified into the three classes belonging to *mellifera* subfamilies. A phylogenetic analysis of MLEs in the three classes revealed two types of clusters, one of which consist of multiple *Apis* MLEs and the other of only *A. mellifera* MLEs. The results obtained provided the "whole picture" of MLEs in *Apis* species. Furthermore, an analysis of long MLEs (>1 kbp or encoding full-length transposases) from the two types of clusters indicated that only one long MLE encoding a complete transposase settled in *Apis* species and burst in the *A. mellifera* genome, while the other long MLEs invaded the *A. mellifera* genome by horizontal transfer after *A. mellifera* speciation and burst in the *A. mellifera* genome, which may also induce the bursting of "settled" MLEs. To the best of our knowledge, this is the first study to provide evolutionary insights into MLEs between multiple related species.

Introduction

Mariner is a class II transposable element (TE) that moves within a host genome in a "copy and paste style" manner (Oosumi et al. 1995; Robertson 1995; Robertson and Lampe 1995; Plasterk et al. 1999). *Mariner* was initially discovered in the *Drosophila mauritiana* genome (Jacobson et al. 1986). The length of the first detected mariner (MOS1) was 1,286 bp.

TE similar to *Mariner* were subsequently identified in the genomes of many plants, vertebrates, and invertebrates, and named mariner-like elements (MLEs). The sizes of MLEs are approximately the same as that of Mos1, but vary copy to copy. The copy numbers of MLEs also vary among species. For example, the copy number of MLEs is estimated to be approximately 14,000 in the human genome (Lander et al. 2001), more than 32,000 in *Arabidopsis thaliana* (Quesneville 2020), and less than 2,000 in *Apis mellifera* (see below). *Mariner* and MLEs encode a transposase that possesses a D,D34D motif, which is characteristic of *mariner* and MLEs (Robertson 1995; Plasterk et al. 1999). Another *mariner*-specific characteristic is the terminal inverted repeat (TIR) and dinucleotide tandem repeats located at both ends of *mariner* that contain transposase-binding sites.

MLEs have been detected using PCR with specific primers designed from the consensus region of the transposase. Detected MLEs have been classified into their subfamilies (e.g. *irritans*, *mellifera*, *mauritiana*, *cecropia*, *and capitata*) (Robertson 1993; Robertson 1995). Previous studies using phylogenetic analyses and *in vitro* transposase experiments indicated the distribution of *mariner* and MLEs in the genomes of many insect species via horizontal transfer (Robertson 1993; Lampe et al. 1996). Evolutionarily, the invading element is considered to have exploded in the host's genome (burst) with the accumulation of mutations. However, MLEs with several mutations may be moved by transposases encoded by other MLEs. The accumulation of more mutations was shown to gradually immobilize MLEs (Witherspoon and Robertson 2003). For example, most of the MLEs detected to date are considered to be non-autonomous in *Drosophila* genomes (Wallau et al. 2014).

PCR-based methods cannot identify all *mariner* and MLEs in genomes because this approach cannot detect some MLEs that have lost their consensus region. Due to the availability of whole genome sequence data and the lower cost of genome sequencing, the methods employed to identify TEs, including *mariner* and MLEs, have been modified. One method involves the use of “RepeatMasker” (Smit A.F.A., Hubley R. & Green P. RepeatMasker Open

4.0. 2013–2015, http://www.repeatmasker.org), which searches sequences in whole genome data that are similar to the consensus TE sequences registered in a sequence library (e.g. Repbase) (Jurka et al. 2005). All types of TEs and autonomous and non-autonomous elements in the whole genome region may be detected using this method.

The genus *Apis* (honey bee species) consists of four major and several minor species, and the whole genome data of the following 4 major species are currently available (three *Apis* species and three *A. cerana* subspecies): *A. mellifera* (Am) (Wallberg et al. 2019), *A. cerana japonica* (Acj) (Yokoi et al. 2018), *A. cerana cerana* Korea native (Ack) (Park et al. 2015), *A. cerana cerana* China native (Acc) (Diao et al. 2018), *A. dorsata* (Ad) (Oppenheim et al. 2020), and *A. florea* (Af) (Table 1). *Apis* genomes contain small numbers of TEs, which mainly consist of class II TEs, particularly MLEs, while the other representative insect genomes (e.g. *Bombyx mori* (Kawamoto et al. 2019), *Aedes aegypti* (Matthews et al. 2018), and *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2008)) contain higher numbers of TEs and MLEs. Therefore, MLE dynamics in *Apis* genomes may be elucidated through analyses of these genomes and detailed individualized data on detected MLEs. Since they have a small number of copies, *Apis* are suitable insects in MLE evolutionary studies.

Difficulties have been associated with the classification of TEs because of variations in their sequences, as described above. Using whole genome information, six *mariner* consensus sequences (Ammers 1-6) have been reported for *A. mellifera* (Weinstock et al. 2006). A more detailed MLE classification was recently developed for *Drosophila* species by Wallau et al. (2014). They re-classified MLEs in *Drosophila* species into 36 classes (Dromars), and also provided detailed insights into Dromars based on the findings of the classified MLEs. Their classification of Dromar was consistent with the MLE subfamilies suggested by Robertson (1993). Therefore, we considered this classification to be more sophisticated, logically accurate, and suitable for application to MLEs in other insect species. In the present study, we extracted Dromar sequences in each Dromar class and built the profile files of each Dromar class. *Apis*

MLEs detected by RepeatMasker in the six *Apis* species genomes were then classified using the profiles built, and a phylogenetic analysis of the *Apis* MLE of each Dromar class was performed (Fig. 1). The results obtained provide a “whole picture” of MLEs in *Apis* species. Moreover, the dynamics of long *Apis* MLEs, defined as >1 kbp or encoding a complete transposase in *Apis* genomes, were revealed using a sequence analysis. Through these results, we provide novel evolutionary insights into *Apis* MLEs.

Results

Mariner-like elements in six *Apis* species

To extract *Apis* MLEs for analysis, RepeatMasker was conducted on the Am, Ack, Acc, Acj, Af, and Ad genomes (Table 1). The output files obtained are shown in Supporting data 1 and 2. Hereafter, the regions detected and annotated as “Mariner” or “Mariner-like element” by RepeatMasker were considered to be MLEs, which may be homologous to the MLE or Mariner sequences in the Repbase library. Approximately 50 to 150 MLEs were detected in the Ack, Acc, Acj, Af, and Ad genomes, while 2,147 MLEs were identified in the Am genome. The detailed and sequence data of the MLEs detected are shown in Supporting data 3 and 4, respectively. To establish whether the identification of many MLEs was *A. mellifera* species- or subspecies-specific, the detection of MLEs by RepeatMasker was performed under the same conditions with the following *A. mellifera* subspecies: *A. mellifera carnica* (Amcar), *A. mellifera caucasica* (Amcau), *A. mellifera intermissa* (Ami), and *A. mellifera mellifera* (Amm) (Table 1). The RepeatMasker results of these *A. mellifera* were not markedly different from those of Am (Supporting data 1 and 2). The numbers of MLEs detected in Amcar, Amcau, Ami, and Amm were 2,247, 2,215, 1,738, and 2,237, respectively, which were not markedly different from those in Am. Therefore, we performed further analyses of Am genome data. While processing the present study, new and high-quality Acc genome data (Acc_new) were published (Wang et al. 2020). To assess whether the results of RepeatMasker, particularly MLEs, were affected by the quality of input genome data, the detection of MLEs by RepeatMasker was conducted under the same conditions using Acc_new. The results of RepeatMasker were not markedly different between Acc and Acc_new (Supporting data 1 and 2). The numbers of MLEs in Acc and Acc_new were 147 and 200, respectively. Therefore, we performed further analyses of Acc genome data. Collectively, the results obtained on MLE numbers in these *Apis* genomes suggested that a small number of MLEs existed in *Apis* genomes, except for the Am genome, and some of these MLEs may burst within the Am genome in the process of evolution.

Building profile of classified Drosophila Mariner-like elements

To elucidate the "whole picture" of MLEs in *Apis* genomes, we attempted to classify the detected MLEs of each *Apis* species. To build profiling data on 36 Dromar groups, we extracted Dromar sequences from 20 *Drosophila* genome sequences and the genomic position data for Dromars reported by Wallau et al. (2014). We only obtained 1,418 Dromar sequences from 3,685 Dromar copies because some *Drosophila* genome data were not available in the public database or the descriptions of some scaffolds in *Drosophila* genome data differed from those in positional data (Supporting data 5). Therefore, sequences for Dromar 3, 7, 18, 19, and 20 were not extracted. Consequently, the Hidden Markov Model (HMM) profiles of 31 Dromars for the classification of *Apis* MLEs, which were generated by the nhmmer program (DNA homology search with profile HMM), were built from the alignment results of each Dromar group (separated and merged alignment and nhmmer profiling data are shown in Supporting data 6 and 7, respectively).

Apis Mariner-like element classification

Apis MLEs extracted by RepeatMasker were classified using the Dromar profiles built. The results of the classification are shown in Fig. 2, and detailed classification results are provided in Supporting data 8. In total, 864 out of approximately 2,000 Am MLEs were mainly classified into Dromar 14, Dromar 17, and Dromar 29, except for one into Dromar 6, all of which were *melliferra* subfamilies (Fig. 2). Most of the MLEs of the five other species that were classified into Dromar were also categorized into Dromar 14, Dromar 17, and Dromar 29 (Fig. 2). A list of all classified MLEs is shown in Supporting data 8. On the other hand, more than 50% of *Apis* MLEs were not classified into any Dromar classes because their E-values were not sufficiently low.

Phylogenetic analysis of *Apis* Marinar-like element sequences.

Sequence alignments were performed using the Dromar and *Apis* MLE sequences belonging to each Dromar class (newick files for phylogenetic trees are shown in Supporting data 9), and a phylogenetic tree of Dromar 14, Dromar 17, and Dromar 29 was constructed based on alignment results because the majority of MLEs were classified into the three classes (Fig. 3), and an original phylogenetic tree figure is shown in supporting data 10. In Dromar 14, there were clearly separated clades in the phylogenetic tree (Fig. 3A). Two of these clades (clades I and III) consisted of only Am MLEs, while the other clade included six *Apis* MLEs plus Dromars (clade II). In Dromar 17, there were four distinct clades (Fig. 3B), one of which included only Am_MLEs (clade I). Clades II and IV included six *Apis* MLEs, while clade III included only Dromars. In the case of Dromar 29, there were five distinct clades (Fig. 3C). Clade I consisted of only Am_MLEs, while clade II included Ad_MLEs and an Ad_MLE. Clade III included six *Apis* MLEs, while clade IV consisted of Dromars plus one Am_MLE. Collectively, these results showed that *Apis* MLEs consisted of three Dromar classes, and the MLEs in each class formed several clusters, one of which consisted of only Am_MLEs and the others of the MLEs of multiple *Apis* species.

Searching for long Mariner-like element

We provided the "whole picture" of the *Apis* MLEs detected. To obtain more detailed insights into the evolution of MLEs across *Apis* species, "long" MLEs were analyzed. Long MLEs are considered to maintain the potential mobile features of MLEs because they possess a *transposase* region. Therefore, we initially listed *Apis* MLEs larger than 1 kbp (hereafter defined as long MLEs) (Table 2 and Supporting data 11). The majority of long MLEs were Am_MLEs, and long MLEs were not identified in the Af genome. Furthermore, long Am_MLEs were equally distributed in all chromosomes, except for the LG_8 chromosome (Fig. 4).

To assess the dynamism of these long *Apis* MLEs, which were not mutated and encoded complete or nearly-complete transposases within or among the six *Apis* genomes, we obtained predicted transposase amino acid sequences using TransDecorder (no long Ad_MLE-encoded transposase was found) (Supporting data 12). A phylogenetic tree was constructed using transposase sequences (Fig. 5). There were three clades. One included several Am_MLEs and single Acc_, Ack_, and Acj_MLEs (IDs were Acc_MLE_100, Acj_MLE_148, and Ack_MLE_6), while the other two included only Am_MLEs, suggesting that the MLEs of Am_, Acc_, Ack_, and Acj_MLEs in this clade settled in *Apis* genomes. To confirm this settlement, we focused our analyses on these MLEs.

Tracking for long Mariner-like elements encoding full transposases between *Apis* species

We searched long MLEs in the clade including multiple *Apis* MLEs that encode full-length transposases (sequences of more than 300 amino acids). We found Acc_MLE_100.p1, Acj_MLE_148.p1, and Ack_MLE_6.p1 in this clade and the sequences of these MLEs were identical, except for one residue, with 99% identity to "*Apis cerana* Mariner transposase" in NCBI GenPept (Accession ID: BAB86288.1) (data not shown) (Sumitani et al. 2002).

To clarify whether this transposase also existed in the Am, Ad, and, Af genomes, we performed a BLAST search (tblastn) against the Am, Ad, and Af genomes using the Acc_MLE_100.p1 sequence as a query. The tblastn results of Am showed that 73 regions with an alignment length > 300 and e-value < 1e-100 were hits (supporting data 14) and the e-value of one of the 73 regions (sequence ID:NC_037644.1, linkage group LG7, range: 12509588-12510547) was 0.0, which included the Am_MLE_972 region and may be a "settling MLE" from *Apis* species in the Am genome. The BLAST results of Ad and Af showed that a single sequence was a hit in each genome (supporting data 14). No MLE coding full-length transposase was identified because MLEs with a length of < 1 kbp could not be detected as "long MLEs." To investigate whether MLEs eventually "settled" in the *Apis* genomes, we performed sequence alignments using these full-length transposase-coding regions plus the outer flanking 500-bp regions of both the 5' and 3' sides (Supporting data 15). As shown in Fig. 6, the TIRs reported by Sumitani et al. (2002) were identified in all six *Apis* species, and most parts of the sequences of transposase-coding regions (underlined in green) plus these TIRs were similar between *Apis* species. Furthermore, although several indels were detected between the *A. cerana* species (Acc, Ack, and Acj) and Am, Ad, and Af in TIRs flanking outer regions, most parts of the sequences in the regions were conserved between the species, suggesting that the MLE eventually settled in this region in these *Apis* species.

Analysis of long Mariner-like element encoding full transposases within the *Apis mellifera* genome

We searched for long MLEs that encode full-length transposases (more than 300 amino acid residues) in the two clades, which consisted of only Am_MLEs (Fig. 5). We did not find any transposase with a sequence of more than 300 amino acids in the two clades, and the longest transposase was Am_MLE_1186.p1 (271 aa) in the smaller clade. We performed a blastp search using Am_MLE_1186.p1 as a query against NCBI-nr. The top hit sequence was “Mariner Mos1 transposase [Stegodyphus mimosarum]” (Accession ID: KFM57872.1), the percent identity, e-value, and query cover of which were 51.16%, 1e-96, and 94%, respectively (data not shown). Using Am_MLE_1186.p1 as a query, we performed a tblastn search against the six *Apis* genome with an E-value < 1e-100. We did not obtain hits in any genomes, except for the Am genome (Supporting data 16). Among the hit sites in the Am genome, there were 5 sites plus one query itself that encoded more than 250 aa. We found two full-length transposases, Am_MLE_182.p1 and Am_MLE_284.p1, in the larger clade. Am_MLE_182.p1 was annotated as Camar1 transposase (Accession ID: AAO12862.1) by blastp against NCBI-nr, the percent identity, e-value, and query cover of which were 48.09%, 1e-120, and 99%, respectively (data not shown). Very similar annotation results to Am_MLE_182.p1 were obtained for Am_MLE_284.p1. Therefore, we performed a tblastn search against the six *Apis* genomes using the two Am_MLEs. Am_MLE_284.p1 hit one sequence with an alignment length > 300 and e-value < 1e-100 in each Af, Acj, and Ack genome, but did not hit any sequence in the Acc and Ad genomes, and the hit sites in Af, Acj, and Ack were the same as those in the Acc_MLE_100 tblastn hit site previously shown. However, the percent identity of these hits was approximately 45% (data not shown). In the Am genome, seven sequences were more than 300 amino acids in length and had 80% identity (Supporting data 17). In the case of Am_MLE_182.p1, there was no tblastn hit site against the Ack, Acj, Acc, Ad, or Af genome (data not shown). In the Am genome, there were seven with a sequence length of more than 300 amino acids and 80% identity. Hit sites were same as the hits of Am_MLE_284.p1 (Supporting data 19). These results suggest that there were two types of MLEs, the Am_MLE_1186 and Am_MLE_284 types, which were only found in the Am genome.

Discussion

In the present study, we searched for MLEs in six *Apis* genomes using RepeatMasker. The number of Am_MLEs was markedly higher than those of other *Apis* species MLEs. MLEs were classified into Dromar classes according to HMM profiles made by Dromar sequences (Wallau et al. 2014). Classification results revealed that the majority of *Apis* MLEs were classified into the Dromar 14, 17, and 19 classes, which are *mellifera* subfamilies in the old classification. The phylogenetic tree of each class showed that these MLEs formed two types of clusters, one of which included only Am_MLEs, while the other included the MLEs of six or multiple *Apis* species. Moreover, there was one long MLE encoding a complete transposase. The single MLE was in the Ad, Af, Acc, Ack, and Acj genomes, and multiple copies were detected in the Am genome. On the other hand, two other types of long MLEs encoding transposases were only found in the *Am* genome.

More Am_MLEs were detected than the five other *Apis* MLEs. Differences in the MLE numbers detected between these *Apis* species were not due to the reference genome qualities of the six *Apis* species. For example, new Acc genome data were published during the processing of the present study (Wang et al. 2020), and we attempted to detect MLEs using these data under the same conditions as those for other *Apis* genomes. Although the contig N50 values of new Acc genome data were more than 150-fold higher than that of the Acc genome used in the present study (Table 1), the numbers of MLEs detected were not markedly different between the two sets of genome data. Moreover, we detected MLEs in multiple Am genomes, including undomesticated Am subspecies, the contig N50 values of which varied. Approximately 1700 to 2,200 MLEs were detected in these Am genomes. Based on these results, differences in MLE profiles between non-Am genomes and the Am genome were not due to technical differences related to nucleotide sequencing or the domestication of Am, but to the evolutionary history of each species. The origin of Am is in the Middle East or North Eastern Africa, while those of *A. cerana*, Ad, and Af are in Asia (Cridland et al. 2017). Based on these findings, the most rational insight related to the different numbers of MLEs between Am and non_Am genomes is that some MLEs existed through *Apis* speciation bursts in the Am genome that were not due to domestication, but to unknown biological factors after *A. mellifera* speciation.

The initial study on the Am genome detected approximately 1,100 MLEs in the genome, and these MLEs were classified into six classes, *AmMar1-6*, which included members of the *mellifera*, *irritans*, and *rosa* subfamilies (Weinstock et al. 2006). In the present study, we detected approximately 2,000 MLEs from the latest version of the Am genome (Wallberg et al. 2019), and classified the MLEs detected based on the findings of Wallatu et al. (2014), who reported more classes than Roberthon et al. (1993). As a result of this classification, all 864 MLEs of the detected MLEs with E-values < 1e-5, except for one, were classified into Dromar 14, 17, and 29, which correspond to *mellifera* subfamilies. These results suggest that the differences observed in classifications between the present and previous studies were due to the detection and classification method, nhmmer. nhmmer adopts a HMM model that detects previously unrecognized sequence features (Wheeler and Eddy 2013). nhmmer, hmmer for nucleotides, has the capacity to detect remote homologs with high sensitivity. Therefore, it is considered to be suitable for analyzing TEs, which are often mutated and highly variable. RepeatMasker also adopted nhmmer and searched TEs with Repbase library. On the other hand, the Dromar classification is considered to be sophisticated because it involves multiple methods and large amounts of genome data. We attempted to perform a new classification method for *Apis* MLEs that combines nhmmer (HMM) and Dromar sequences, and provided detailed classifications and novel insights into *Apis* MLEs. Therefore, our classification method using nhmmer and the MLE sequences of the other species may be employed in a TE analysis. Based on our classification in the present study, we classified the MLEs of the five other *Apis* species. The majority of MLEs of Acc, Ack, Acj, Ad, and Af were classified into Dromar 14, 17, and 29, while a few were classified into *mauritiana* subfamilies. Collectively, our classifications showed that MLEs belonging to Dromar 14, 17, and 29 in the Acc, Ack, Acj, Ad, and Af genomes increased their copies in the Am genomes.

Phylogenetic trees of 1kbp < MLEs encoding long transposases revealed that there were two types of clusters: one consisting of multiple *Apis* species MLEs and the other of only Am_MLEs. The single former type of MLE, which encode complete transposases previously identified in Acj (Sumitani et al. 2002), was found in the Acc, Ack, Acj, Ad, and Af genomes and MLEs were placed in a locus in each genome, while many copies of these MLEs, including one MLE, placed in the locus were found in the Am genome. A number of short-length MLEs were also detected in non-*Apis* genomes. Further detailed analyses demonstrated that these MLEs invaded the genome via horizontal transfer, moved before *Apis* species speciation, settled in *Apis* species, and burst after Am speciation, implying that some factors allow MLEs to burst in Am (Fig. 7). On the other hand, the latter type of Am_MLEs, which carry a similar transposase to Camar1 transposase [*Chymomyza amoena*] or Mariner Mos1 transposase [*Stegodyphus mimosarum*], were only found in the Am genome. These results showed that these types of MLEs invaded the Am genome by horizontal transfer and burst after Am speciation. The reason why MLEs burst after Am speciation remains unknown. Moreover, the transposases encoded by these types of MLEs may make the former type of MLEs “move” in the Am genome. The Himar1 transposase is involved in MLE possessing *in vitro* (Lampe et al. 1996). Based on the present results and previous findings, the most plausible insight is that invaded MLEs burst and the transposases of MLEs also allow “settled” MLEs to burst in the Am genome (resulting in a loss of mobility); however, the former type of MLE encoding a full-length transposase possessed the ability to move “itself” (Fig. 7).

In conclusion, we herein detected MLEs from the genome data of six *Apis* species, performed nhmmer-based *Apis* MLE classifications and phylogenetic analyses, and conducted a detailed analysis of MLEs encoding full-length transposases. The number of Am_MLEs detected in Am genomes, including multiple Am subspecies, was approximately 50-100-fold higher than those of the other *Apis* MLEs, which were not due to domestication. The majority of *Apis* MLEs were classified into three Dromar classes, which are *mellifera* subfamilies. Long

MLEs with lengths >1 kbp were divided into two types. Further detailed analyses provided the plausible evolutionary insight that one type of MLE settled in the *Apis* genome and burst in the Am genome, while the others invaded and burst in the Am genome, thereby allowing the other "settled" MLEs to also burst in this genome (Fig. 7). To the best of our knowledge, this is the first study to provide evolutionary insights into MLEs between multiple related species.

Materials and Methods

MLE detection and extraction of sequence data

The six *Apis* genomes plus other versions and several *A. mellifera* subspecies sequences used in the present study are shown in Table 1.

RepeatMasker (Smit A.F.A., Hubley R. & Green P. RepeatMasker Open 4.0. 2013–2015, http://www.repeatmasker.org) was used to detect TEs in the six *Apis* genomes, as described in our previous study (Yokoi et al. 2018). RepeatMasker output (.tbl and .out) files are shown in Supporting data 1 and 2, respectively. Data on MLE were extracted from outfiles, and a single ID was allocated to each detected element (Supporting data 3). To extract each mariner sequence, positional data in Supporting data 3 and BLAST software were utilized (Supporting data 4) (Altschul et al. 1990).

Classification of *Apis* MLEs

To build the hmmer profiles of Dromars, Dromar sequences were extracted using the *Drosophila* genome data and Dromar location data reported by Wallau et al. (2014) (Supporting data 5). However, some Dromar sequences were not extracted because genome sequence IDs in *Drosophila* genome data differed from those described in Dromar location data. Therefore, several Dromars were not used in the present study. The sequences of each Dromar were aligned by Clustal Omega (Clustal Omega - 1.2.4) (Sievers et al. 2011), and the hmmer profiles of Dromars were made by hmmer-building (HMMER 3.2.1) (Supporting data 6) (Wheeler and Eddy 2013). Using the nhmmer search program with merged Dromar hmm profiles (Supporting data 7), *Apis* MLEs similar to each Dromar profile were identified (E-values < 1e-5) (Wheeler and Eddy 2013). The output data and table results of nhmmer are shown in Supporting data 18 and 19, respectively. Based on hmmer search results, each MLE was classified into hit Dromar. If a MLE was hit to multiple Dromars, it was classified into Dromar with the lowest e-value.

Phylogenetic analysis

Dromars and classified MLE were aligned by Clustal Omega-1.2.4, an approximately-maximum-likelihood phylogenetic analysis was performed using FastTree version 2.1.10 (original data are shown in Supporting data 8) (Price et al. 2009), and phylogenetic trees were drawn in MEGAX version 10.1.8-1 and iTOL version 5 (Kumar et al. 2018; Letunic and Bork 2019).

Long MLE analysis

The distribution of Am_MLEs > 1 kbp in *A. mellifera* chromosomes was visualized by ChromoMap (Anand 2019). To obtain the predicted transposases of long MLEs, TransDecorder version 5.5.0 was used with default settings (https://github.com/TransDecoder/TransDecoder/wiki). EMBOSS needle (6.6.0.0) was employed to compare the sequences of long MLEs encoding complete transposases plus flanking sites. Using these sequences, multiple sequence alignments were performed by Clustal Omega-1.2.4 (Sievers et al. 2011), and the results of the alignments were visualized by Boxshade Server with default settings (https://embnet.vital-it.ch/software/BOX_form.html).

Availability of data

Analyzed data and additional results have been uploaded to figshare as Supporting data.

Supporting data 1

Out files of RepeatMasker from six *Apis* genomes plus *A. mellifera* subspecies and new *Apis cerana cerana* China native genome data. DOI:10.6084/m9.figshare.11914278

Supporting data 2

Tbl files of RepeatMasker from six *Apis* genomes plus *A. mellifera* subspecies and new *Apis cerana cerana* China native genome data. DOI: 10.6084/m9.figshare.11984451

Supporting data 3

MLE data on six *Apis* species plus *A. mellifera* subspecies and new *Apis cerana cerana* China native genome data. DOI: 10.6084/m9.figshare.11914263

Supporting data 4

MLE sequences. DOI: 10.6084/m9.figshare.11913309

Supporting data 5

Dromar sequences (Wallau et al. 2014) used in the present study. DOI:10.6084/m9.figshare.11956350

Supporting data 6

Alignment and hmmer profile files of Dromars. DOI:10.6084/m9.figshare.11962614

Supporting data 7

Merged Dromar hmmer profile for MLE classifications. DOI:10.6084/m9.figshare.11953734

Supporting data 8

List of *Apis* MLEs and classification results DOI:10.6084/m9.figshare.12045189

Supporting data 9

Original data of the phylogenetic analysis in newick files. DOI:10.6084/m9.figshare.11955621

Supporting data 10

Original phylogenetic tree figures of Dromar 14 (DOI:10.6084/m9.figshare.13235168), Dromar 17 (DOI:10.6084/m9.figshare.12798179), and Dromar 29 (DOI:10.6084/m9.figshare.12798263).

Supporting data 11

ID list of all long *Apis* MLEs. DOI: 10.6084/m9.figshare.13235168

Supporting data 12

Predicted amino acid sequence of long MLE transposase. DOI: 10.6084/m9.figshare.13237385

Supporting data 13

Original figures of phylogenetic trees for transposases of *Apis* long MLEs. DOI: 10.6084/m9.figshare.13237388

Supporting data 14

tblastn results of the Acc_MLE_100 transposase as a query against the Am genome with e-value < 1e-100. DOI: 0.6084/m9.figshare.13237397

Supporting data 15

Sequence data of transposases plus outer flanking 500-bp regions of TIRs of 5’ and 3’ sides of *Apis* MLEs encoding full-length transposases (all.fa.gz) and alignment result data of Clustal omega (all.aln.gz). DOI: 10.6084/m9.figshare.14094587

Supporting data 16

Tblastn results of Am_MLE_1186 against the *A. mellifera* genome. DOI: 10.6084/m9.figshare.13238258

Supporting data 17

Tblastn results of Am_MLE_182.p1 and Am_MLE_284 against the *A. mellifera* genome. DOI: 10.6084/m9.figshare.13238246

Supporting data 18

Results of nhmmer of *Apis* MLEs against Dromar profiles. DOI: 10.6084/m9.figshare.11953746

Supporting data 19

Result tables of nhmmer of *Apis* MLEs against Dromar profiles. DOI: 10.6084/m9.figshare.11953752

Acknowledgments

This work was supported by ROIS-DS-JOINT (026RP2019 and 030RP2018) to KY and by the National Bioscience Database Center of the Japan Science and Technology Agency (JST) and

Hiroshima Prefectural Government to HB. We express our thanks to Drs. Masatsugu Hatakeyama and Yuudai Masuoka for their valuable discussions.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403–410.

Anand L. 2019. chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes. *bioRxiv*:605600.

Cridland JM, Tsutsui ND, Ramírez SR. 2017. The complex demographic history and evolutionary origin of the western honey bee, Apis mellifera. *Genome Biology and Evolution* 9:457–472.

Diao Q, Sun L, Zheng Huajun, Zeng Z, Wang S, Xu S, Zheng Huoqing, Chen Y, Shi Y, Wang Y, et al. 2018. Genomic and transcriptomic analysis of the Asian honeybee Apis cerana provides novel insights into honeybee biology. *Scientific Reports* [Internet] 8:1–14. Available from: https://www.nature.com/articles/s41598-017-17338-6

Jacobson JW, Medhora MM, Hartl DL. 1986. Molecular structure of a somatically unstable transposable element in Drosophila. *Proceedings of the National Academy of Sciences* 83:8684–8688.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* 110:462–467.

Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, Fujiyama A, Kiuchi T, Yamamoto K, Shimada T. 2019. High-quality genome assembly of the silkworm, Bombyx mori. *Insect Biochemistry and Molecular Biology* 107:53–62.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* 35:1547–1549.

Lampe DJ, Churchill ME, Robertson HM. 1996. A purified mariner transposase is sufficient to mediate transposition in vitro. *The EMBO journal* 15:5470–5479.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. 2001. Initial sequencing and analysis of the human genome.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research* 47:W256–W259.

Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, Glassford WJ, Herre M, Redmond SN, Rose NH. 2018. Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature* 563:501–507.

Oosumi T, Belknap WR, Garlick B. 1995. Mariner transposons in humans. *Nature* 378:672–672.

Oppenheim S, Cao X, Rueppel O, Krongdang S, Phokasem P, DeSalle R, Goodwin S, Xing J, Chantawannakul P, Rosenfeld JA. 2020. Whole Genome Sequencing and Assembly of the Asian Honey Bee Apis dorsata. *Genome Biology and Evolution* 12:3677–3683.

Park D, Jung JW, Choi B-S, Jayakodi M, Lee J, Lim J, Yu Y, Choi Y-S, Lee M-L, Park Y. 2015. Uncovering the novel characteristics of Asian honey bee, Apis cerana, by whole genome sequencing. *BMC genomics* 16:1.

Plasterk RH, Izsvák Z, Ivics Z. 1999. Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends in genetics* 15:326–332.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol Biol Evol* [Internet] 26:1641–1650. Available from: https://academic.oup.com/mbe/article/26/7/1641/1128976

Quesneville H. 2020. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. *Mobile DNA* 11:1–13.

Robertson HM. 1993. The mariner transposable element is widespread in insects. *Nature* 362:241–245.

Robertson HM. 1995. The Tcl-mariner superfamily of transposons in animals. *Journal of insect physiology* 41:99–105.

Robertson HM, Lampe DJ. 1995. Distribution of transposable elements in arthropods. *Annual review of entomology* 40:333–357.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.

Sumitani M, Lee JM, Hatakeyama M, Oishi K. 2002. Cloning and characterization of Acmar1, a mariner-like element in the asiatic honey bee, Apis cerana japonica (Hymenoptera, Apocrita). *Archives of Insect Biochemistry and Physiology: Published in Collaboration with the Entomological Society of America* 50:183–190.

Tribolium Genome Sequencing Consortium, Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, et al. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature* 452:949–955.

Wallau GL, Capy P, Loreto E, Hua-Van A. 2014. Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. *BMC genomics* 15:727.

Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. *BMC genomics* 20:275.

Wang Z-L, Zhu Y-Q, Yan Q, Yan W-Y, Zheng H-J, Zeng Z-J. 2020. A Chromosome-Scale Assembly of the Asian Honeybee Apis cerana Genome. *Frontiers in Genetics* 11:279.

Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson GE, Worley KC, Evans JD, Maleszka R, Robertson HM, et al. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*

[Internet] 443:931–949. Available from: https://www.nature.com/articles/nature05260

Wheeler TJ, Eddy SR. 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* [Internet] 29:2487–2489. Available from: https://academic.oup.com/bioinformatics/article/29/19/2487/186765

Witherspoon DJ, Robertson HM. 2003. Neutral evolution of ten types of mariner transposons in the genomes of Caenorhabditis elegans and Caenorhabditis briggsae. *Journal of molecular evolution* 56:751–769.

Yokoi K, Uchiyama H, Wakamiya T, Yoshiyama M, Takahashi J-I, Nomura T, Furukawa T, Yajima S, Kimura K. 2018. The draft genome sequence of the Japanese honey bee, Apis cerana japonica (Hymenoptera: Apidae). *European Journal of Entomology* 115.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Author contributions

Conceived and designed the experiments: K.Y., K.K., and H.B.

Analyzed the data: K.Y., K.K., and H.B.

Contributed to the writing of the manuscript: K.Y. and H.B.

All authors discussed the data and contributed to manuscript preparation. K.Y. supervised the project.

All authors read and approved the final manuscript.

Competing interests

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Figure legends

Fig. 1 Scheme of data analyses in the present study

Fig. 2 *Apis* MLEs allocated to Dromars

The number of *Apis* MLEs allocated to each Dromar (e-value <1e-5). The left drawing shows the molecular phylogeny of the genus *Apis* based on the study by Lo et al. (2010). The color of each column indicates the subfamilies of *mariner* to which Dromar belongs, based on Wallau et al. (2014). Unnumbered Dromars are those whose sequences were not extracted from the genome. Dromars with white numbers filled in black comprised more than 50 MLEs. Asterisks indicate the Dromars consisting of MLEs from more than two *Apis* species. Detailed classification lists are shown in Supporting data 8.

Fig. 3 Phylogenetic trees of three Dromar classes.

Phylogenetic trees based on the alignments of *Apis* MLEs of the three Dromar classes (Dromar 14 (A), 17 (B), and 29(C)), to which more than 50 MLEs are classified. Orange-, red-, and black-curved lines in these figures indicate clusters including only Am_MLEs, multiple *Apis* species MLEs, and mainly Dromars. Roman numerals in the trees indicate independent clusters. Bootstrap values symbolized the size of circles. The phylogenetic trees of other classes are shown in supporting data 10. These original figures downloaded from iTOL are shown in supporting data 10.

Fig. 4 Distributions of long Am_MLEs. LG_1 to LG_16 indicate the chromosome names of *A. mellifera*.

Yellow markers indicate long Am_MLEs. Bold yellow markers indicate two MLEs placed at very close positions to each other. The figure was drawn by chromoMap

Fig. 5 Phylogenetic tree of transposases of *Apis* long MLEs

Phylogenetic tree of predicted transposases of *Apis* long MLEs. Orange- and red-curved lines in these figures indicate clusters including only Am_MLEs and multiple *Apis* species MLEs. Original figures downloaded from iTOL are shown in supporting data 13.

Fig. 6 Sequence alignment results of MLEs coding full-length transposases plus flanking 500-bp regions of 5’ and 3’ sides

The alignment data of MLEs coding full-length transposases plus flanking 500-bp regions of the 5’ and 3’ sides (Supporting data 15) were visualized by the Boxshade server. Nucleotide sites that are preserved or similar between multiple *Apis* species are shaded in black or gray. The TIR sites reported by Sumitani et al. (2002) are shown in red boxes, and transposase-coding regions are underlined in green.

Fig. 7 Schematic summary of evolutionary insights into *Apis* MLEs revealed in the present study.

Tables

Table 1 *Apis* genome assemblies used in the present study

Organism name	**GenBank assembly accession**	**Contig N50**	**Abbreviation in this study**
A. mellifera	GCA_003254395.2	5,382,476	Am
A. cerana japonica	GCA_002217905.1	179,487	Acj
A. cerana cerana Korea native	GCA_001442555.1	43,751	Ack
A. cerana cerana China native	GCA_002290385.1	21,160	Acc
A. dorsata	GCA_009792835.1	30,868	Ad
A. florea	GCA_000184785.2	24,915	Af

A. mellifera carnica (Carniolan honeybee)	GCA_013841245.1	2,692,667	Amcar
A. mellifera intermissa	GCA_000819425.1	504	Ami
A. mellifera caucasica (Caucasian honeybee*)*	GCA_013841205.1	3,303,520	Amcau
A. mellifera mellifera (German honeybee)	GCA_003314205.1,	5,131,172	Amm
A. cerana cerana China native (new version)	GCA_011100585.1	3,898,192	Acc_new

Table 2 Numbers of long MLEs (1 kbp < length)

Organism name	**Number of long MLEs**	**Total number of MLEs**
A. mellifera	80	2,145
A. cerana japonica	7	158
A. cerana cerana Korea native	6	140
A. cerana cerana China native	7	145
A. dorsata	3	163
A. florea	0	163

Figures

Fig. 1

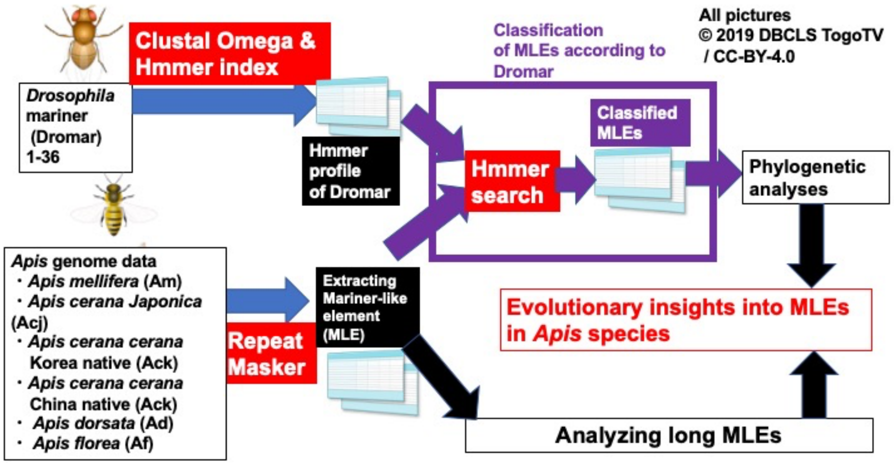
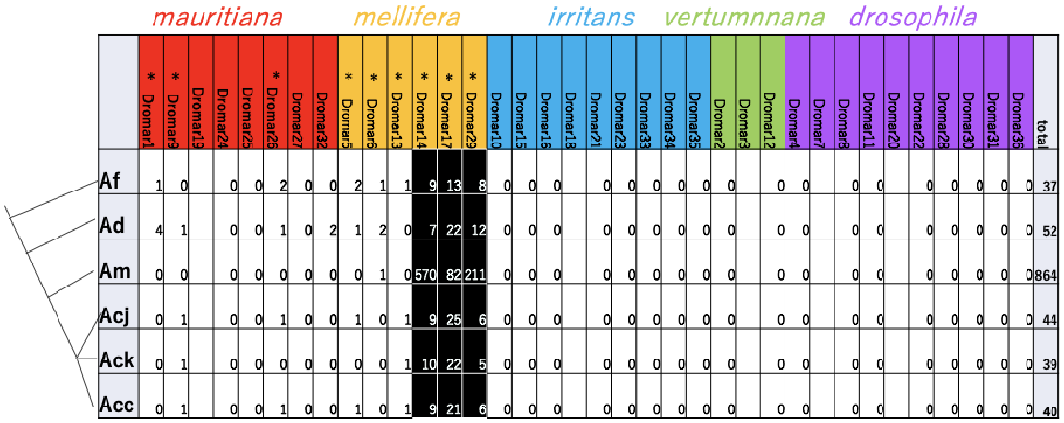

Clustal Omega & Hmmer index
Drosophila mariner (Dromar) 1-36
Hmmer profile of Dromar
Classification of MLEs according to Dromar
All pictures © 2019 DBCLS TogoTV / CC-BY-4.0
Classified MLEs
Hmmer search
Phylogenetic analyses
Apis genome data
・Apis mellifera (Am)
・Apis cerana Japonica (Acj)
・Apis cerana cerana Korea native (Ack)
・Apis cerana cerana China native (Ack)
・Apis dorsata (Ad)
・Apis florea (Af)
Repeat Masker
Extracting Mariner-like element (MLE)
Evolutionary insights into MLEs in Apis species
Analyzing long MLEs

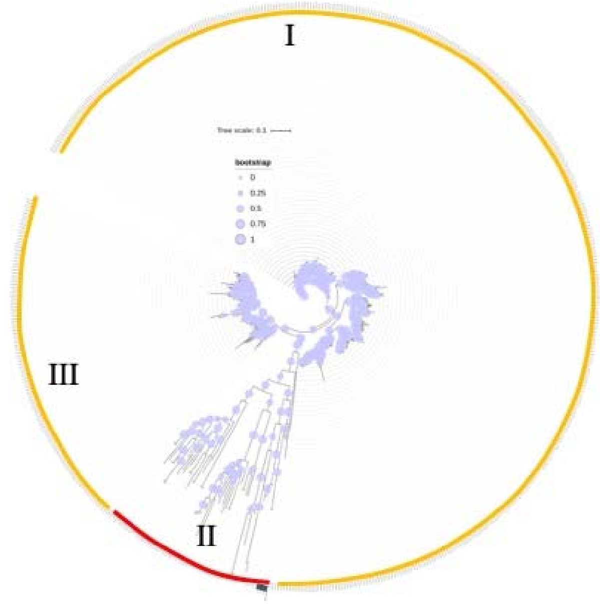
Fig. 2



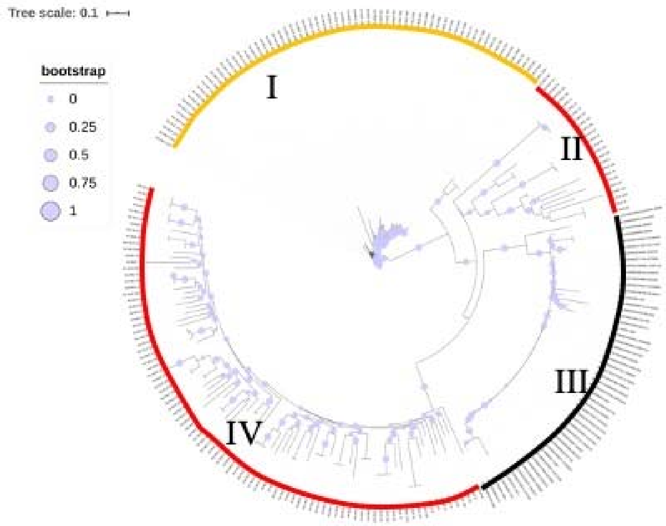
	mauritiana								*mellifera*						*irritans*									*vertumnnana*			*drosophila*										
	* Dromar1	* Dromar9	Dromar19	Dromar24	Dromar25	* Dromar26	Dromar27	Dromar32	* Dromar5	* Dromar6	* Dromar13	* Dromar14	* Dromar17	* Dromar29	Dromar10	Dromar15	Dromar16	Dromar18	Dromar21	Dromar23	Dromar33	Dromar34	Dromar35	Dromar2	Dromar3	Dromar12	Dromar4	Dromar7	Dromar8	Dromar11	Dromar20	Dromar22	Dromar28	Dromar30	Dromar31	Dromar36	total
Af	1	0		0	0	2	0	0	2	1	1	9	13	8	0	0	0		0	0	0	0	0	0		0	0		0	0		0	0	0	0	0	37
Ad	4	1		0	0	1	0	2	1	2	0	7	22	12	0	0	0		0	0	0	0	0	0		0	0		0	0		0	0	0	0	0	52
Am	0	0		0	0	0	0	0	0	1	0	570	82	211	0	0	0		0	0	0	0	0	0		0	0		0	0		0	0	0	0	0	864
Acj	0	1		0	0	1	0	0	1	0	1	9	25	6	0	0	0		0	0	0	0	0	0		0	0		0	0		0	0	0	0	0	44
Ack	0	1		0	0	0	0	0	0	0	1	10	22	5	0	0	0		0	0	0	0	0	0		0	0		0	0		0	0	0	0	0	39
Acc	0	1		0	0	1	0	0	1	0	1	9	21	6	0	0	0		0	0	0	0	0	0		0	0		0	0		0	0	0	0	0	40

*Consisting of MLEs from more than two *Apis* species

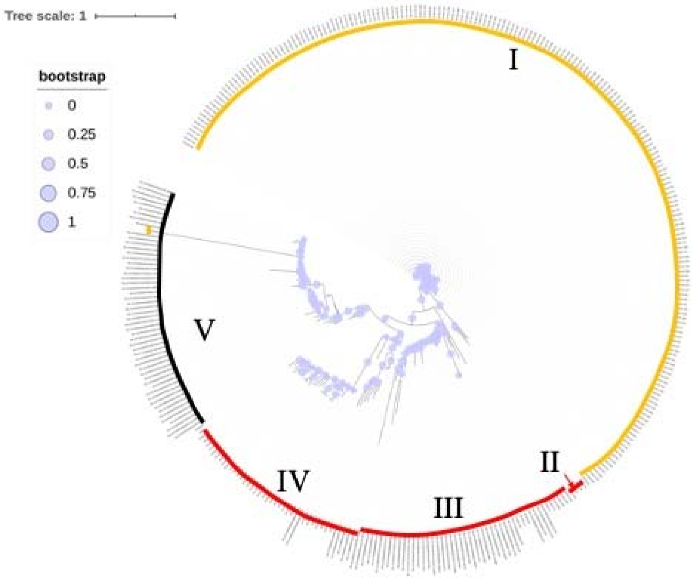
Fig. 3 (A)


I
Tree scale: 0.1
bootstrap
0
0.25
0.5
0.75
1
III
II

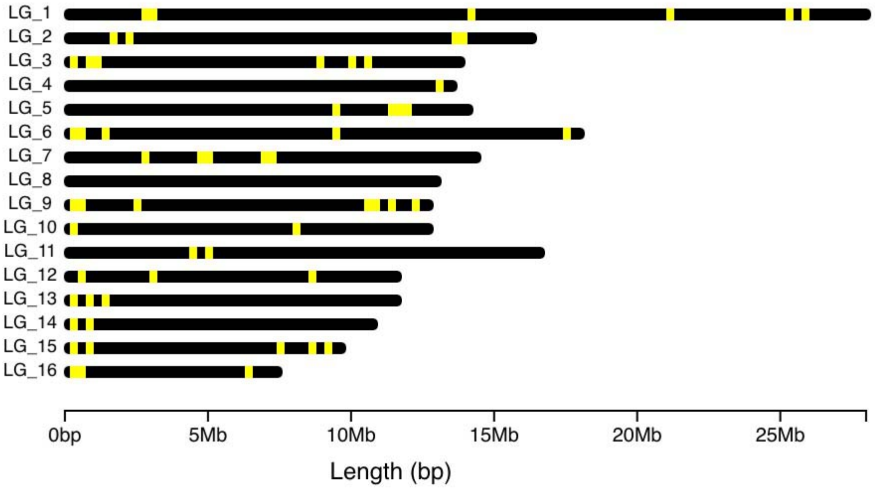
(B)


Tree scale: 0.1
bootstrap
0
0.25
0.5
0.75
1
I
II
III
IV

(C)


Tree scale: 1
bootstrap
0
0.25
0.5
0.75
1
I
II
III
IV
V

616 **Fig. 4**



617

618 **Fig. 5**

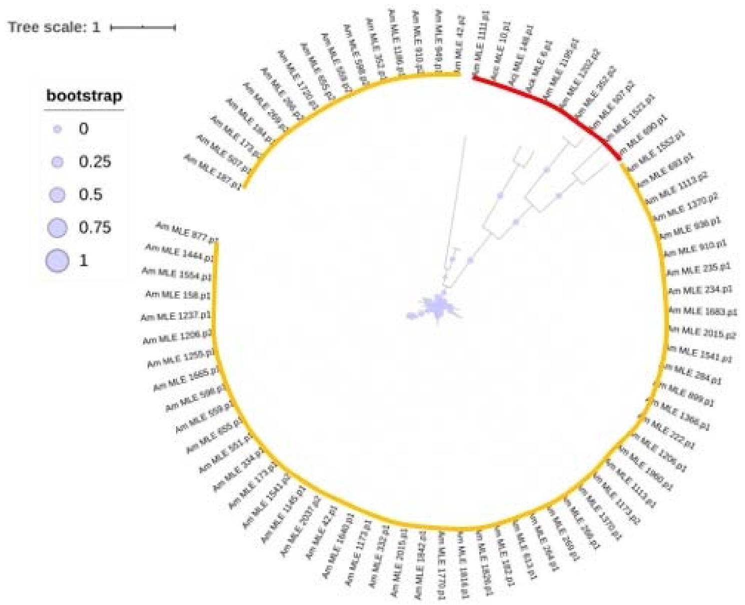
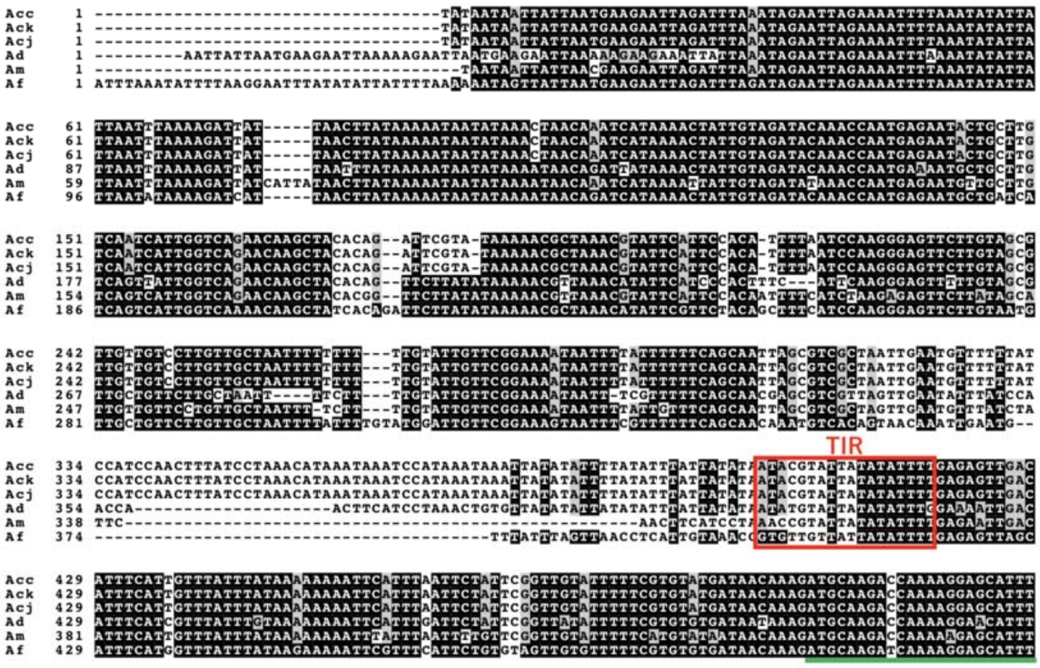
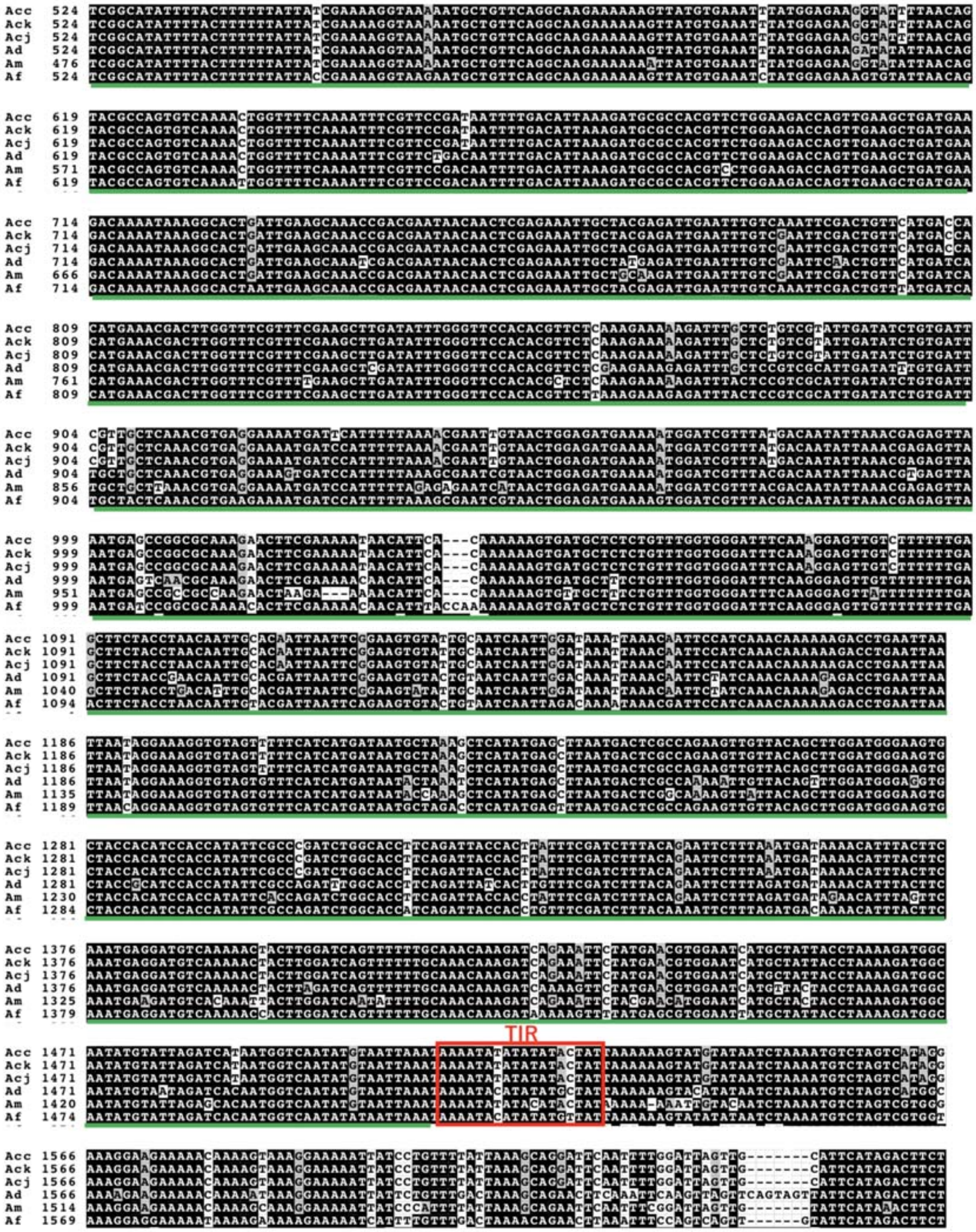

Tree scale: 1
bootstrap
0
0.25
0.5
0.75
1

Fig. 6


Acc
Ack
Acj
Ad
Am
Af
TIR


Acc
Ack
Acj
Ad
Am
Af
TIR

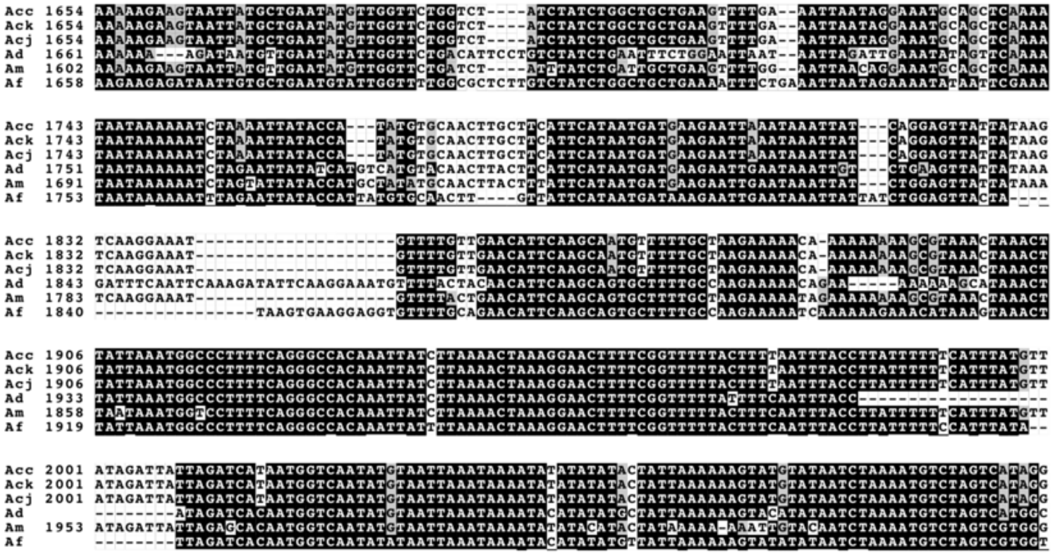
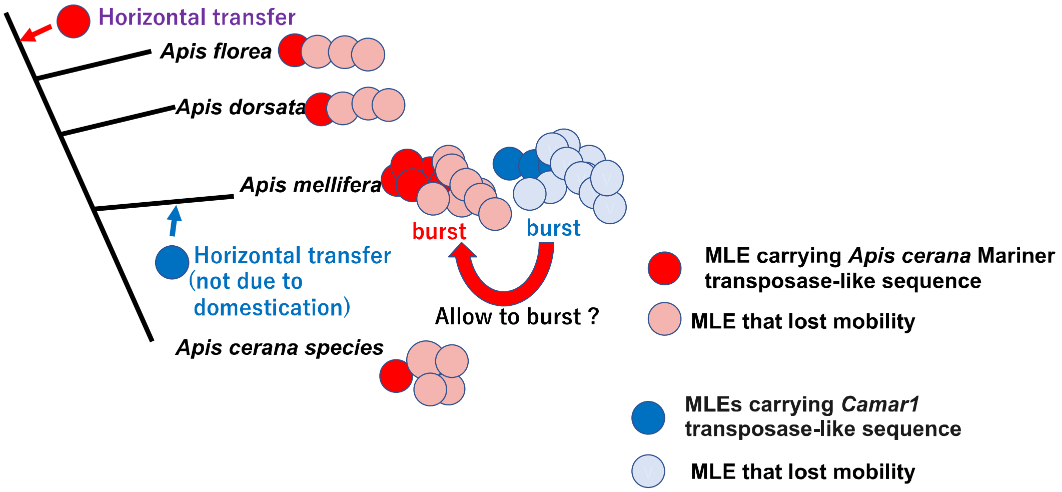

Acc 1654
Ack 1654
Acj 1654
Ad 1661
Am 1602
Af 1658
Acc 1743
Ack 1743
Acj 1743
Ad 1751
Am 1691
Af 1753
Acc 1832
Ack 1832
Acj 1832
Ad 1843
Am 1783
Af 1840
Acc 1906
Ack 1906
Acj 1906
Ad 1933
Am 1858
Af 1919
Acc 2001
Ack 2001
Acj 2001
Ad
Am 1953
Af

Fig. 7


Horizontal transfer
Apis florea
Apis dorsata
Apis mellifera
Horizontal transfer
(not due to
domestication)
burst
burst
Allow to burst ?
Apis cerana species
MLE carrying *Apis cerana* Mariner transposase-like sequence
MLE that lost mobility
MLEs carrying *Camar1* transposase-like sequence
MLE that lost mobility