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: Kakeru Yokoi

E-mail addresses:

Kakeru Yokoi: yokoi123@affrc.go.jp

Kiyoshi Kimura: kimura@affrc.go.jp

Hidemasa Bono: bonohu@hiroshima-u.ac.jp

Abstract

Background: *Mariner* and mariner-like elements (MLEs) are distributed in various species and their sequences are highly diverse. In previous reports, a few transposable element in the genomes of *Apis* species mainly consist of *mariner* and MLE. For further insight of *Apis* MLEs, detailed classifications of *Apis* MLE and sequences analysis of long MLEs, which may potentially encode the transposase, are needed.

Results: Much more MLEs were detected in *A. mellifera* genome compared to other *Apis* species genomes. They were classified into 31 *Drosophila* MLE classes. In this classification, almost all of MLEs were classified into the three classes belonging to *mellifera* subfamilies, suggesting that *Apis* MLEs which exist thorough *Apis* species derived from single MLE belonging to *mellifera* subfamily. Phylogenetic analysis using MLEs in the three classes showed that there two types of clusters, of which one consist of multiple *Apis* species MLEs, and others of only *A. mellifera* MLEs. Long MLEs analysis showed that only one long MLE encoding complete transposase was found in all *Apis* genome except for *A. mellifera* genome, and the MLE and multiple MLEs similar to it were found in *A. mellifera* genome. The analysis also showed that other several long MLEs encoding complete transposase were found only in *A. mellifera* genome.

Conclusions: Almost all of *Apis* MLEs are *mellifera* subfamilies. Of these MLEs, one types of them settled in *Apis* species and burst in *A. mellifera* genome. The other one of MLEs invaded into *A. mellifera* genome by horizontal transfer and exploded in *A. mellifera* genome. This is the first example of the finer aspects of MLE evolution among closely related species,

Key words

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

Background

Mariner is a class II transposable element (TE) that moves within a host genome in a "copy and paste style" manner [1–3]. *Mariner* was initially discovered in the *Drosophila mauritiana* genome [4]. The length of first detected mariner (MOS1) is 1286 bp. TE similar to *Mariner* were subsequently identified in the genomes of many plants, vertebrates, and invertebrates, and named mariner-like elements (MLEs). The size of MLEs are about the same as one of Mos1, but varies copy to copy. The copy number of MLEs varies from species to species as well. For example, a copies of MLEs in the human genome is estimated about 14,000 [5] and one of in *Apis mellifera* is less than 2,000 copies (see below). *Mariner* and MLEs encode a transposase that possesses a D,D34D motif, which is characteristic of *mariner* and MLEs [3]. Another *mariner*-specific characteristic is the terminal inverted repeat (TIR) located at both ends of *mariner* that contains transposase-binding sites.

MLE has been detected using PCR with specific primers designed form consensus region of the transposase. Detected MLEs have been are classified their subfamilies (e.g. *irritans, mellifera, mauritana, cecropia, and capitata*)[2,6,7]. Previous studies indicated that *mariner* and MLE distributed in the genomes of many species via horizontal transfer [2]. Evolutionary, the invading element is thought to have exploded in the host's genome (burst) and with the accumulation of mutations, the MLEs becomes immobilized gradually. Most of the detected MLEs are thought to be non-autonomous.

The PCR-based methods cannot detect all *mariner* and MLEs in genomes because a MLE that has lost the consensus region cannot be found in this way. These days, due to the availability of whole genome sequence data and the lower cost of genome sequencing, the methods employed to identify TE, including *mariner* and MLE, 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 consensus TE sequences registered in a sequence library (e.g. Repbase) [8]. This method detects all types of TE and autonomous and non-autonomous elements in the whole genome region, thereby tracking MLEs and providing further insights into the horizontal transfer of MLEs between host species.

The genus *Apis* (honey bee species) consists of four major and several minor species, and the following whole genome data of major 4 species are currently available (three *Apis* species and three *Apis cerana* subspecies): *A. mellifera* (Am) [9], *A. cerana japonica* (Acj) [10], *A. cerana cerana* Korea native (Ack) [11], *A. cerana cerana* China native (Acc) [12], *A. dorsata* (Ad) [13], and *A. florea* (Af) (Table 1). *Apis* genomes contain small numbers of TE,

which mainly consist of class II TE, particularly MLEs [10,14]. Thus, through analysis using these whole genomes, and detected MLEs data, MLEs dynamics in *Apis* genomes could be revealed. Being a small number of copies, in MLE evolutionary studies, *Apis* are one of the suitable insects.

Difficulties have been associated with the classification of TE 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* [14]. However, this classification is just the results of RepeatMasker therefore it is not well considered.
Recently, using Drosophila genome, Wallau et al., (2014) re-classified the MLEs in *Drosophila* species into 36 classes (Dromars) [15]. This classification is more sophisticated logically correct. The old PCR methods did not follow evolutionary back ground. In the present study, we extracted Dromar sequences in each Dromar class and built the profile files of each Dromar classes by Clustal Omega and hmmer programs with the Dromar sequences. *Apis* MLEs were then classified using these profiles by nhmmer (DNA homology search with profile Hidden Markov Models). Moreover, a phylogenetic analysis of the *Apis* MLE of each Dromar class was performed (Fig. 1). The dynamics of long *Apis* MLEs, defined as larger than 1kbp in *Apis* genomes, were revealed by using sequence analysis. Through these results, we provide novel evolutionary insights into *Apis* MLEs.

Results

MLEs in six *Apis* species

RepeatMasker was used to detect MLEs in the Am, Ack, Acc, Acj, Af, and Ad genomes. The output files obtained are shown in Supporting data 1 and 2. Approximately 50 to 150 MLEs were detected in the Ack, Acc, Acj, Af, and Ad genomes, while 2147 MLEs were identified in the Am genome. 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. mellifer*a species- or subspecies-specific, the detection of MLEs by RepeatMasker with the same conditions was performed with the *A. mellifera* subspecies, these are *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 Am (Supporting data 1 and 2), and the numbers of MLEs in Amcar, Amcau, Ami and Amm were 2247, 2215, 1738 and 2237, respectively. The numbers of MLEs in these

genomes were not markedly different from those in Am. Therefore, we decided to perform further analyses against Am genome data. While processing the present study, new and high-quality Acc genome data (Acc_new) were published [16]. To clarify whether the results of RepeatMasker particularly in MLEs, were changed, the detection of MLEs by RepeatMasker with the same conditions was performed. The results of RepeatMasker were not markedly different between Acc and Acc_new (Supporting data 1 and 2), and the number of MLEs in Acc and Acc_new were 147 and 200, respectively. Therefore, we decided to perform further analyses of Acc genome data. Collectively, the results of MLE numbers in these *Apis* genomes suggested that a small number of MLEs existed in *Apis* genomes, except for Am genome, and some of these MLEs might burst within the Am genome in the process of evolution.

Dromar profiling

To build profiling data on 36 Dromar groups, we extracted Dromar sequences using 20 *Drosophila* genome sequences and the positional data of Dromar reported by Wallau et al. (2014). We only obtained 1418 Dromar sequences from 3685 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 nhmmer profiles of 31 Dromars for the classification of *Apis* MLEs 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 MLE classification

Apis MLEs were classified using the extracted *Apis* MLE sequences and Dromar profiles. 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 2000 Am MLEs were classified into only mainly Dromar 14, Dromar 17 and Dromar 29 except for one into Dromar6, which are all *melliferra* subfamilies, while no MLE were classified into the other subfamilies. (Fig. 2). The MLEs of the other five species classified into *mauritiana* and *mellifera* subfamilies, and most of MLEs were classified into Dromar 14, Dromar 17 and Dromar 29. A list of all classified MLEs is shown in Supporting data 8. The results suggested that MLEs in Apis genome mainly consist of *melliferra* subfamilies.

Phylogenetic analysis of *Apis* MLE and Dromar 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 tree figure is shown in supporting data 10. In Dromar 14, there were distinct 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 them included only Am_MLEs (clade I). Clades II and IV include six *Apis* MLEs while clade III included 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 two Ad_MLEs and an Ad_MLE. Clade III included the six *Apis* MLEs while clade IV consisted of Dromars plus one Am_MLE. These results suggested that there are two groups of Apis MLEs. We hypothesized that some MLEs forming the Am_MLE-only group in the phylogenetic trees may invade to *A. mellifera* genomes by horizontal transfer and burst within the *A. mellifera* genome, while the other MLEs forming six *Apis* MLEs settle in *Apis* genomes.

Long MLE analysis

MLEs extracted using RepeatMasker include collasped, non-functional and short "junk" MLEs. Further elucidating the evolution insight of MLE in Apis, "long" MLE were analyzed. The long MLEs are thought to maintain the potential mobile feature of MLE. To confirm this hypothesis, we initially listed *Apis* MLEs larger than 1 kbp (long MLEs) (Table 4 and Supporting data 11). The majority of long MLEs are Am_MLEs. Long Am_MLEs were equally distributed in all chromosomes, except for the LG_8 chromosome (Fig. 4), suggesting that the burst of long Am_MLEs is not a chromosomal-local but across these all chromosomes.

To assess the dynamism of these long *Apis* MLEs within or among the six *Apis* genomes, we obtained the predicted transposase amino acid sequences using TransDecorder (Supporting data 12). A phylogenetic tree was constructed by using the transposase sequences (Fig. 5). There were three clades. One included Am_, Acc_, Ack_ and Acj_MLEs while the other two included only Am_MLEs, suggesting that MLEs consisting of Am_, Acc_, Ack_, and Acj_MLEs settled in Apis genomes, while those consisting of Am_MLEs invaded and burst in *A. mellifera* genome.

To confirm the settlement, we searched long MLEs in the clade including multiple *Apis* MLEs that code full-length transposases (sequence 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 three sequences were the same, except for one residue, with 99% identity of "Apis cerana Mariner transposase" in NCBI-nr (Accession ID: BAB86288.1) (data not shown) [12]. To clarify whether the transposase also existed in Am, Ad and, Af genomes, we performed BLAST search (tblastn) against the Am, Ad and Af genomes. The tblastn results of Am showed that 73 MLEs with an alignment length > 300 and e-value < 1e-100 were hits (supplemental data 14) and the E_value of one of the 73 MLEs (sequence ID:NC_037644.1 range:12509588-12510547) was 0.0, which may have been a "settling MLE" from Apis species in the Am genome. The tblastn results of Ad and Af showed that a single sequence was a hit in each genome. (supplemental data 14). To investigate whether the Acc_MLE_10 settled in the *Apis* genomes, we compared the nucleotide sequences of the Acc_MLE_10 plus the flanking 500bp on both sides with the hit MLEs plus flanking 500bp on both sides in Am, Ack, Acj, Ad, Af (supporting data 15). More than 80% identities were observed in all combinations and the internal sites of the two sequences were mostly same while most outer sites in both sides were different, showing that the TIR of each MLE was conserved. Collectively, these results suggested that the MLEs settled through *Apis* genomes and burst in the Am genome.

We searched long MLEs in the two clades including only Am_MLEs that code full-length transposases (sequence of more than 300 amino acids) (Fig. 5). Of the two clades, we found no transposase with a sequence of more than 300 amino acids and the longest transposase was Am_MLE_1186.p1 (271 aa) in the smaller clade. We performed a blastp search by using Am_MLE_1186.p1 as 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, we performed a tblastn search against the six *Apis* genome with 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). In the case of Am_MLE_182.p1, the same annotation results of Am_MLE_284.p1 were obtained. Consequently, 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 the Af, Acj and Ack were the same as the Acc_MLE_100 tblastn hit site as previously reported. However, the percent of identities of these hits were 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. The hit sites were same as the hits of Am_MLE_284.p1 (Supporting data 17). These results suggested that there were two types of MLEs, Am_MLE_1186 and Am_MLE_284 types, which were only found in the Am genome and burst within it.

Discussion

In the present study, we extracted MLEs from six *Apis* genomes. The number of Am_MLEs was markedly higher than those of the other *Apis* species MLEs. MLEs were classified into Dromar classes according to Hidden Markov Model (HMM) profiles made by Dromar sequences [15]. Classification results revealed that the majority of Apis MLEs were classified into *mellifera* subfamilies in old classification. The phylogenetic tree of each class showed that these MLEs formed two types of clusters, of which one included only Am_MLEs, while the other included the MLEs of six or multiple *Apis* species. Moreover, there was one long MLE-encoding complete transposase, which might settle in the Ad, Af, Acc, Ack, and Acj genomes and might explode in the Am genome, and two other types of long Am_MLEs that only existed in the *A. mellifera* genome and might explode within it.

More Am MLEs were detected than the other five *Apis* MLEs. Differences in the MLE numbers detected were not due to the reference genome qualities of the six *Apis* species. We detected MLEs by RepeatMasker using multiple *Apis* genome data (Table 3). For example, new Acc genome data were published during the processing of the present study [16], and we performed MLE detection using these data. 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, the numbers of MLEs detected were not markedly different between the two sets of genome data. Moreover, we detected MLEs by multiple *A. mellifera* genomes, the contig N50 values of which varied. Approximately 1700 to 2200 MLEs were detected in these Am genomes. Based on these results, differences in MLEs between not Am genome and Am genome were not

due to artificial or technical differences, suggesting that different feature of the MLEs among each *Apis* species seemed to due to evolutionary history of each species.

The initial study on the Am genome detected approximately 1100 MLEs in the genome, and these MLEs were classified into six classes, *AmMar1-6*, which included members of the *mellifera*, *irritans*, and *rosa* subfamilies [14]. In the present study, we detected approximately 2000 MLEs from the latest version of the Am genome [9], and 864 MLEs of the detected MLEs with E-value < 1e-5 were classified. All 864 MLEs were allocated to *mellifera* subfamilies. This difference in classifications between the present and previous studies may be due to the detection and classification method, nhmmer. nhmmer adopted a HMM model that detects previously unrecognized sequence features [17]. In detail, nhmmer, hmmer for nucleotide, can detect remote homologs as sensitively as possible [17]. Thus, we think nhmmer is suitable for analyzing TEs, which are often mutated and highly varied. RepeatMakser also adopted nhmmer and searched TEs with Repbase library. On the other hand, the Dromar classification is considered to be sophisticated because the classification was achieved by multiple methods and many 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, the majority of MLEs of Am, Acc, Ack, Acj, Ad, and Af were classified into *mellifera* subfamilies, while a few were classified into *mauritiana* subfamilies. The present results further supported *Apis* MLEs exploding in the Am genome.

Phylogenetic trees revealed that there were two types of clusters, one of which consisted of only Am_MLEs, and the other of several *Apis* species MLEs. It is plausible that the former type of MLE invaded the genome in horizontal transmission and exploded after the bees diverged in speciation from other species of bees while MLEs of the latter type invaded the genome in horizontal transmission and settled through *Apis* species genomes and explode in the Am genome before the species divergence. This aspect was also consistent with the analysis of the long MLEs, which have putative transposases sequence and may maintain mobility. In addition, the long MLE that encoded the complete transposase, *A. cerana* Mariner transposase, was not found in the Ad, Af, or Am genome. However, the MLE encoding the complete transposase was identified in the Ad, Af, and Am genomes by a tblastn search because these MLEs were detected as a length < 1 kbp by RepeatMasker and were filtered out. One MLE that encoded the same sequence of *A. cerana* Mariner transposase was found in the Am genome, and several MLEs that encoded transposases that were highly similar to *A. cerana* Mariner

transposase were also detected. These results suggested that the MLEs encoding *A. cerana* Mariner transposase invade horizontally through *Apis* evolutions and "burst" in the Am genome, but not in other *Apis* species. On the other hand, other types of Am_MLEs which carry the transposase similar to Camar1 transposase [*Chymomyza amoena*] were only found in the Am genome. These types of MLEs may invade the only Am genome by horizontal transfer and burst within the Am genome. The reason why MLEs only burst in the Am genome remains unknown. The origin of *Apis* honey bee is Asia, except for Am, whose origin is Europe-Africa. On the other hand, in most beekeeping, the Am is used and is widespread in the all over world as domestic insects. These facts may be related to this aspect. More detailed structural analysis of Am and other *Apis* genome must be needed. This is the first example of the finer aspects of MLE evolution among closely related species, and perhaps the first time in other TEs. Integrating all the results, the evolutionary aspect of MLEs belonging to the Dromar is shown in Figure 6.

Conclusions

We detected MLEs from the genome data of six *Apis* species, and performed nhmmer-based *Apis* MLE classification and phylogenetic analyses. About over 50-100 times of Am_MLEs were detected than these of the other *Apis* MLEs. Almost all of Apis MLEs classified were classified into *mellifera* subfamilies. The long MLEs of which length are over 1kbp divided into two types. One type of MLEs settled in Apis genome and burst in Am genome, the other invade into and burst in Am genome. We showed firstly provided evolutional insight *Apis* MLEs (Fig. 6).

Materials and Methods

MLE detection and extraction of sequence data

The six *Apis* genome 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 TE in the six *Apis* genomes, as described in our previous study [10]. 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) [18].

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) [15]. 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) [19], and the hmmer profiles of Dromars were made by hmmer-building (HMMER 3.2.1) (Supporting data 6) [17]. Using the nhmmer search program with merged Dromar hmm profiles (Supporting data 7), the Apis MLEs similar to each Dromar profile were identified (E-values < 1e-5) [17]. 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) [20], and phylogenetic trees were drawn in MEGAX version 10.1.8-1 and iTOL version 5 [21,22].

Long MLE analysis

The distribution of long Am_MLEs in *A. mellifera* chromosomes was visualized by ChromoMap [23]. To obtain the predicted transposase 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 used to compare the sequences of the long MLEs encoding the complete transposase plus flanking site [24].

Data availability

Analyzed data and additional results have been uploaded to figshare as Supporting data. These data are available via the DOI shown in the Supporting data legends.

References

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

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

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

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

5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409,860–921

6. Sumitani M, Lee JM, Hatakeyama M, Oishi K. 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 2002;50:183–190.

7. Yamada K, Kawanishi Y, Yamada A, Tokuda G, Gurung RD, Sasaki T, et al. A novel cluster of mariner-like elements belonging to mellifera subfamily from spiders and insects: implications of recent horizontal transfer on the South-West Islands of Japan. Genetica. 2014;142:149–160.

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

9. Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, et al. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC genomics. 2019;20:275.

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

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

12. Diao Q, Sun L, Zheng H, Zeng Z, Wang S, Xu S, et al. Genomic and transcriptomic analysis of the Asian honeybee Apis cerana provides novel insights into honeybee biology. Scientific Reports 2018;8:1–14.

13. Oppenheim S, Cao X, Rueppel O, Krongdang S, Phokasem P, DeSalle R, et al. Whole Genome Sequencing and Assembly of the Asian Honey Bee Apis dorsata. Genome Biology and Evolution. 2020;12:3677–3683.

14. Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson GE, et al. Insights into social insects from the genome of the honeybee Apis mellifera. Nature 2006;443:931–49.

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

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

17. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics 2013;29:2487–9.

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

19. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7:539.

20. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular biology and evolution. 2009 26:1641–50. A

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

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

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

24. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic acids research. 2015;43:W580–W584.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All supporting data have been uploaded in figshare (DOI:10.6084/m9.figshare.c.4894503)

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 trees figures of Dromar14 (DOI:10.6084/m9.figshare.13235168), Dromar17 (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 MLEs transposase. DOI: 10.6084/m9.figshare.13237385

Supporting data 13

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

Supporting data 14

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

Supporting data 15

The nucleotide sequences data of the Acc_MLE_10 plus flanking 500bp and long MLEs encoding complete transposase plus flanking 500bp in Am, Ack, Acj, Ad, Af, and results of compering of the Acc sequences with the other sequences.
DOI: 10.6084/m9.figshare.13237427

Supporting data 16

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

Supporting data 17

Tblastn results of Am_MLE_182.p1 and Am_MLE_284 against *A. mellifera* genome. DOI:

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

Competing interests

The authors declare no conflict 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.

Funding

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.

Authors' 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.

Figures & Tables

Fig. 1 Scheme of data analyses in the present study

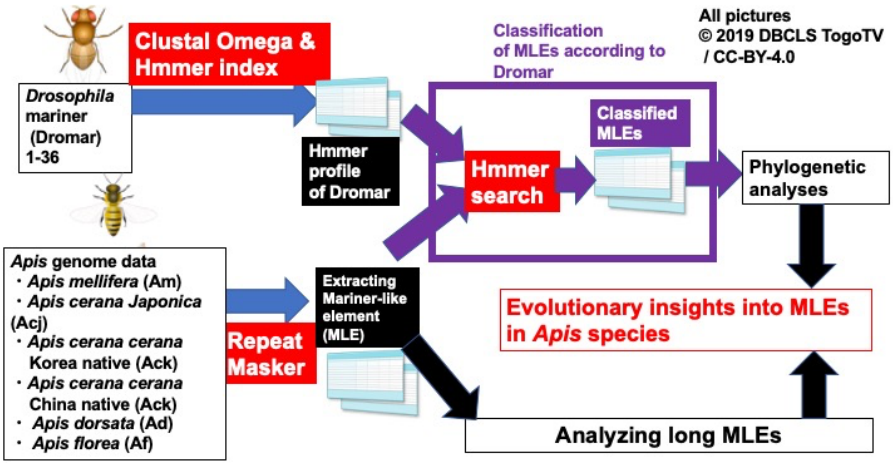
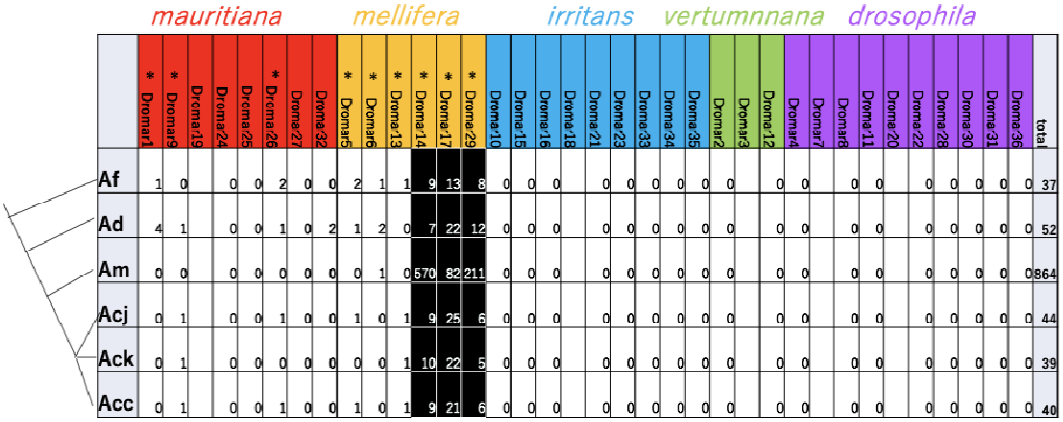


Fig. 2 Apis MLEs allocated to Dromars

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 could not be extracted from the genome. The Dromars with white numbers filled in black comprised more than over 50 MLEs. Asterisks indicate the Dromars consisting of MLEs from more than two *Apis* species. Detailed classification lists are shown in Supporting data 8.



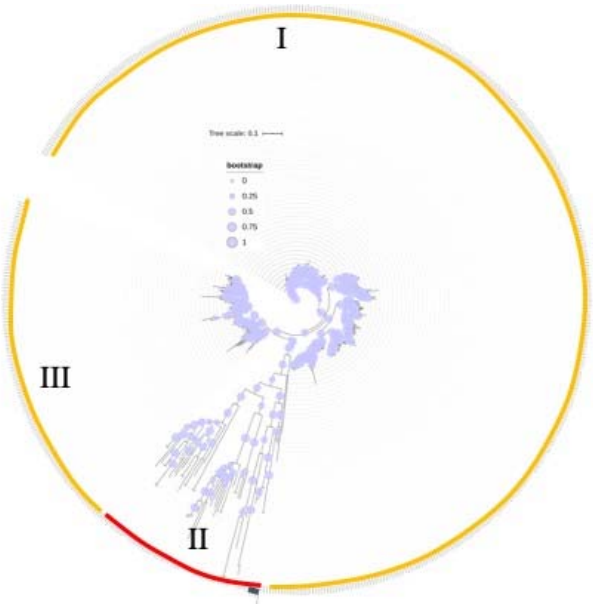
	* Dromar1	* Dromar9	Dromar19	Dromar24	Dromar25	* Dromar26	Dromar27	Dromar32	* Dromar5	* Dromar6	* Dromar13	* Dromar14	* Dromar17	* Dromar29	Dromar10	Dromar15	Dromar16	Dromar18	Dromar21	Dromar23	Dromar33	Dromar34	Dromar35	Dromar2	Dromar3	Dromar12	Dromar4	Dromar7	Dromar6	Dromar11	Dromar20	Dromar22	Dromar28	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

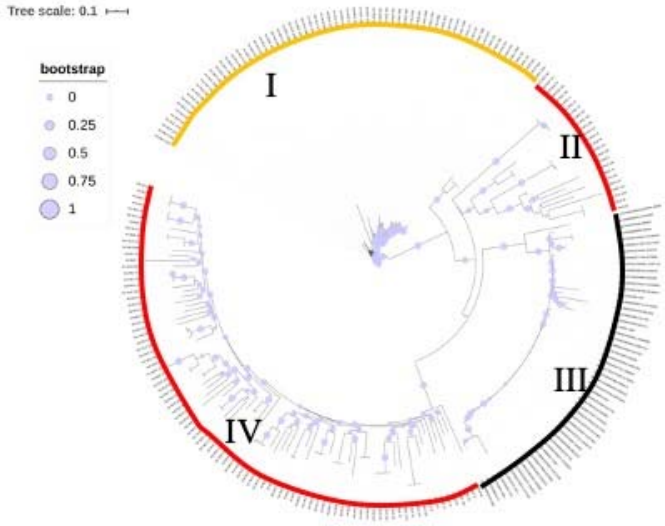
Fig. 3 Phylogenetic trees of three Dromar classes.

Phylogenetic trees based on the alignments of *Apis* MLEs of the three Dromar classes (Dromar 14, 17, and 29), 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.

(A) Dromar 14



(B) Dromar 17


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

(C) Dromar 29

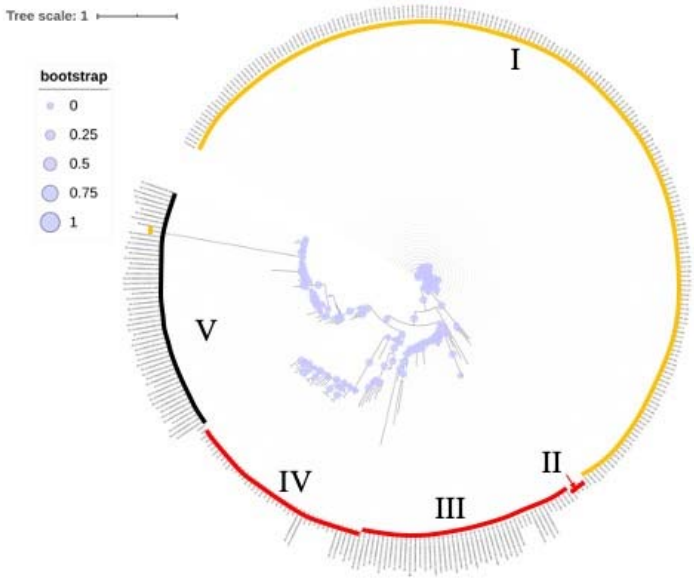

Tree scale: 1
bootstrap
0
0.25
0.5
0.75
1
I
II
III
IV
V

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

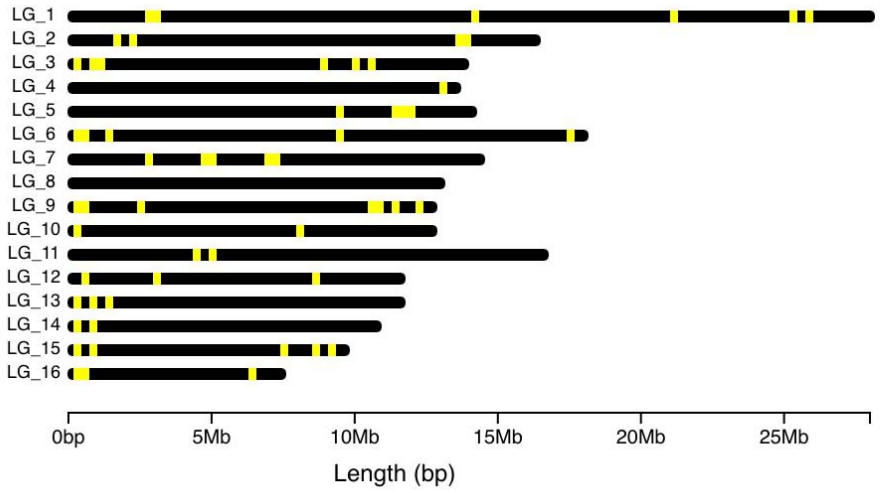


Fig. 5 Phylogenetic tree of transposases of *Apis* long MLEs

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. These original figures downloaded from iTOL are shown in supporting data 13.

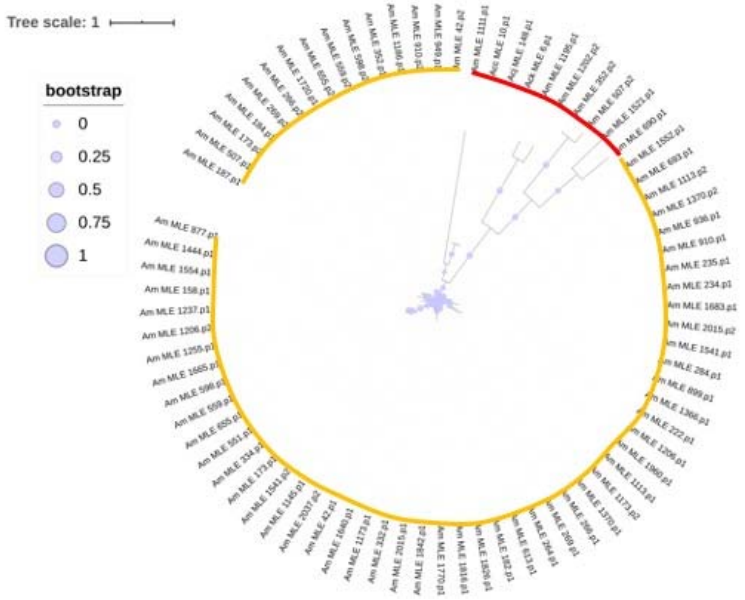


Fig. 6 Evolutional insight of Apis MLEs

Schematic summary of evolutionary insights into *Apis* MLEs revealed by this study.

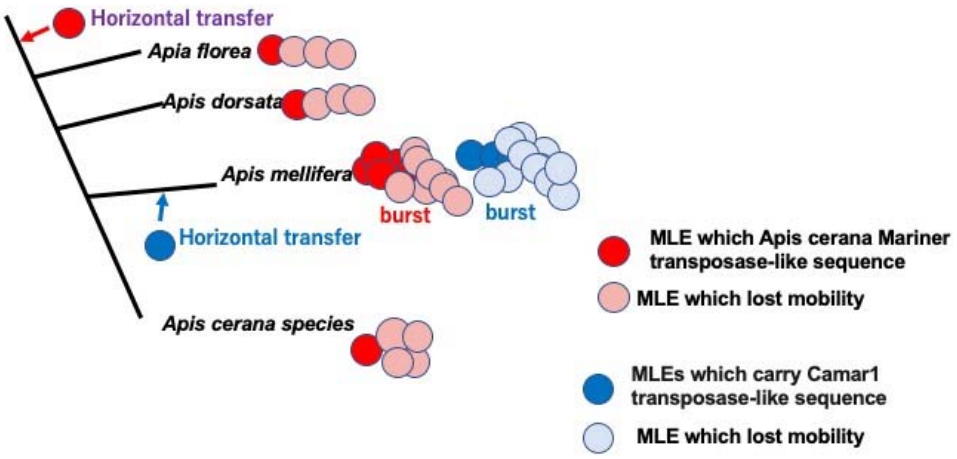


Table 1 *Apis* genome assemblies used in the present study

Organism name	GenBank assembly accession	Contig N50
A. mellifera	GCA_003254395.2	5,382,476
A. cerana japonica	GCA_002217905.1	179,487
A. cerana cerana Korea native	GCA_001442555.1	43,751
A. cerana cerana China native	GCA_002290385.1	21,160
A. dorsata	GCA_009792835.1	30,868
A. florea	GCA_000184785.2	24,915
A. mellifera carnica (Carniolan honeybee)	GCA_013841245.1	2,692,667
A. mellifera intermissa	GCA_000819425.1	504
A. mellifera caucasica (Caucasian honeybee)	GCA_013841205.1	3,303,520
A. mellifera mellifera (German honeybee)	GCA_003314205.1,	5,131,172
A. cerana cerana China native (new version)	GCA_011100585.1	3,898,192

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