scientific reports

OPEN

Discovery of the first Tn630 member and the closest homolog of IS630 from viruses

Yanping Hu[1,7], Guangyou Duan[3,7], Haohao Yan[2], Yutong Guo[2], Jia Chang[2], Mingbing Zhou[4], Shuangyong Yan[5], Wenjing Li[1], Cihan Ruan[6]✉ & Shan Gao[2]✉

IS630/Tc1/*mariner* (ITm) represents the most widely distributed superfamily of DNA transposons in nature. Currently, bioinformatics research on ITm members primarily involves collecting data of existing and emerging members and organizing them into new groups or families. In the present study, our survey revealed that Tc1 and IS630 members have a broad host range, spanning across all six biological kingdoms (bacteria, fungi, plantae, animalia, archaea and protista) and viruses. The primary discoveries include the first Tn630 member—Tn630-NC1 and the closest homolog of IS630 from viruses—Tc1-C#1. By incorporating our discoveries into existing knowledge, we proposed a model to elucidate the formation of composite transposons. Organization of Tc1 and IS630 members into groups across biological kingdoms facilitates data collection for future research, particularly on their horizontal transfer between different kingdoms. The formation of composite transposons may result from asymmetric of terminal inverted repeats. IS630 should be merged with Tc1 into a single family IS630/Tc1. Furthermore, IS630 and its homologs constitute a valuable resource for studying horizontal gene transfer between gut bacteria and phages, opening up new avenues for research in this field.

Keywords TE, ITm, Mariner, Pogo, HGT

IS630/Tc1/*mariner* (ITm) represents the most widely distributed superfamily of DNA transposons in nature[1]. Members of the ITm superfamily are identified by the featured domains of their transposases, which contain the catalytic pockets responsible for cleaving DNA strands. The featured domains of these transposases have active-site motifs that contain three acidic amino-acid (aa) residues DDE or DDD[1]. The ITm superfamily includes four typical families (Tc1, *mariner*, IS630, and *pogo*). The first member of the ITm superfamily, named Tc1 (GenBank: X01005), was discovered in *Caenorhabditis elegans* in 1983[2]. Later, the Tc1 family was defined to include homologs of Tc1 in animals, plants, filamentous fungi and yeast[3]. *Mariner* (GenBank: X78906), IS630 (GenBank: X05955), and *pogo* (GenBank: X59837), as the first members of the *mariner*, IS630, and *pogo* families, were discovered in *Drosophila mauriliana*[4], *Shigella sonnei*[5], and *Drosophila melanogaster*[6], respectively. With more ITm members identified, it was concluded that all Tc1 transposases identified in fungi, invertebrates, and vertebrates contain a DD34E motif in their DDE domains, while most *mariner* transposases identified in flatworm, insects, and vertebrates contain a DD34D motif in their DDD domains[7]. Therefore, the DDxE/D (x represents the number of aa residues between the second D and the third E/D) motif is used as a highly conserved feature for the identification and classification of ITm members.

Understanding origins and evolution of ITm members across diverse organisms poses a challenging and profoundly significant area in basic research. This endeavor necessitates the collection of data from various sources and the development of methodologies to systematically categorize these members into meaningful groups or families. Currently, bioinformatics research on ITm members primarily involves collecting data of existing and emerging members, coupled with the organization of them into new groups or families (e.g., Sailor[8]). This is particularly evident with the appearance of more complete, even full-length genomes[9] through

[1]Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, Qinghai, People's Republic of China. [2]College of Life Sciences, Nankai University, Tianjin 300071, People's Republic of China. [3]School of Life Sciences, Qilu Normal University, Jinan 250200, Shandong, People's Republic of China. [4]The State Key Laboratory of Subtropical Silviculture, Bamboo Industry Institute, Zhejiang A&F University, Hangzhou 311300, Zhejiang, People's Republic of China. [5]Tianjin Academy of Agricultural Sciences, Institute of Crop Research, Tianjin 300381, People's Republic of China. [6]Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95050, USA. [7]These authors contributed equally: Yanping Hu and Guangyou Duan. ✉email: cruan@scu.edu; gao_shan@mail.nankai.edu.cn

advanced sequencing technologies such as PacBio and Nanopore DNA-seq[10]. The first concern addressed by the present study is a significant bias in the collected data, primarily originating from an emphasis on closely related members within some species (e.g., *Drosophila melanogaster*) and a simultaneous oversight of members from others (e.g., viral species). The second concern is that most of reported groups contan ITm members restricted to one or two kingdoms, making it difficult to investigate horizontal transfer (HT) of ITm members across kingdoms. The third concern — how to classify ITm members, is still an open question. Given that huge amounts of latent ITm members in public databases have not been identified, the determination of the number of families under the ITm superfamily remains elusive. Traditionally, researchers classified ITm members based on their DDxE/D motifs. However, the report of new ITm members with diverse motifs (e.g., DD37E, DD37D, DD38E, and DD39D[7]) complicated this simplistic classification. These motifs differed from classical ones such as DD35E for IS630, DD34E for Tc1, DD34D for *mariner*, and DD30D for *pogo*. Two typical examples were IS630-AB1 (AB representing *Acinetobacter baumannii*) and Tc1-OP1 (OP representing *Ogataea parapolymorpha*)[3]. IS630-AB1 (DD34E) deviated from the classical view as an atypical member of the IS630 family, while Tc1-OP1 (DD40E) was the first member of the Tc1 family in yeast[3]. Recent research findings underscored that the DDxE/D motif lacks sufficient conservation to divide the ITm superfamily into families or groups under a family. Additionally, comparatively new families under the ITm superfamily were questioned regarding their acceptance. For instance, Dupeyron et al. proposed that the *pogo* family is more extensive and diverse than previously acknowledged, suggesting it could be defined as a distinct superfamily[11]. Therefore, a comprehensive collection and systematic organization of ITm members are imperative for comprehending the intricate division of the ITm superfamily.

The present study started with a survey of Tc1 and IS630/Tn630 members spanning across biological kingdoms, resulting in primary discoveries including the first Tn630 member — Tn630-NC1 and the closest homolog of IS630 from viruses—Tc1-C#1. Extensive studies of their homologs revealed a broad host range of Tc1 and IS630 members, spanning across six biological kingdoms (bacteria, fungi, plantae, animalia, archaea and protista) and viruses. In addition, we proposed a procedure to organize Tc1 and IS630 members into IS630/Tc1 groups across biological kingdoms, aiming to facilitate data collection for future research, particularly on their HT between different kingdoms. By analyzing the IS630/Tc1 groups constructed by our established procedure, we reached several conclusions.

Results and discussion

Discovery of the first Tn630 member

IS630 and Tc1 are DNA transposons, which have been discovered in *S. sonnei*[5] and in *C. elegans*[2], respectively, and they also represent two distinct families under the ITm superfamily, including homologs from bacterial species (i.e., members from the IS630 family or IS630 members) and non-bacteria species (i.e., members from the Tc1 family or Tc1 memebers), respectively. The present study defined Tn630 as a composite transposon that consists of two IS630 simple transposons as components (IS630 components). Although the acquisition of Tn630 or its homologs (i.e., members from the Tn630 family or Tn630 members) from bacterial genomes is theoretically possible, none of Tn630 members had been reported before the present study and all previously reported IS630 members were not coupled into composite transposons. The present study started with a survey of Tc1 and IS630/Tn630 members, spanning across biological kingdoms using the NCBI NT and NR databases. To facilitate the survey, the name formats were designed as Tc1-XYn and IS630-XYn for Tc1 and IS630 members, respectively, where X and Y are the initial letters of the genus and the species names (# is used to indicate unknown genus or species), respectively and n is a number to distinguish members from the same genus and species. In addition, Tn630 members were named using the name format of Tn630-XYn. For example, three new Tc1 or IS630 members reported in our previous study[3] were named using the above name formats and they are: (1) Tc1-OP1 (OP represents *Ogataea polymorpha*), the first member of the Tc1 family in yeast; (2) Tc1-MP1 (MP represents *Mucor piriformis*), the homolog of Tc1-OP1 in filamentous fungi; and (3) IS630-AB1 (AB represents *Acinetobacter baumannii*), an atypical member of the IS630 family in the classical view. The 5,659-bp reference sequence of Tc1-OP1 consists of a 169-bp 5' terminal inverted repeat (TIR), a 245-bp 5' untranslated region (UTR), a 3,468-bp open reading frame (ORF) of transposase (1,155-aa), a 1,608-bp 3' UTR, and a 169-bp 3' TIR, while the 1,688-bp reference sequence of Tc1-MP1 consists of a 141-bp 5' TIR, a 60-bp 5' UTR, a 1,305-bp ORF (434 aa), a 41-bp 3' UTR, and a 141-bp 3' TIR. Different from Tc1-OP1 and Tc1-MP1 containing intact ORFs, IS630-AB1 was recovered by inserting an adenine (A) residue to its ORF to obtain its 871-bp reference sequence, consisting of a 19-bp 5' UTR, a 849-bp recovered ORF and a 3-bp 3' UTR. In our previous study[3], the homologs of IS630-AB1 had been detected in almost all *Acinetobacter* spp., however, none of them were coupled into Tn630-AB1.

Using the IS630-AB1 transposase for homology search (Methods and materials), a comprehensive identification revealed a minimum of 55 copies of IS630 members, named IS630-NC1 (NC represents *Nitrosomonas communis*), in the genome of *N. communis* strain Nm2 (GenBank: CP011451.1) from the NCBI NT database. Among these 55 copies of IS630-NC1, only five (Supplementary file 1) contained intact ORFs of transposases. Conversely, the remaining 50 copies featured either partial ORFs or full-length ORFs with insertion and deletions (InDels) leading to premature translation termination codons (PTCs). Notably, a significant portion of these InDels manifested as one adenine insertion or two-adenine (AA) deletions within a polyA region (denoted as $[A]_n$, n is the repeating times of A). This observation suggested that PTCs in the transposase ORFs of ITm members are likely predominantly caused by InDels within Short Tandem Repeats (STRs), particularly within polyA regions. Among the five copies containing intact ORFs, at least four were identified as intact transposons, containing complete TIRs and "TATA" at their 5' and 3' ends. Finally, Tn630-NC1 (CP011451: 3800760–3807217) was identified from the 50 copies of IS630-NC1. However, Tn630-NC1 (Supplementary file 1) had degenerated into an inactive status, due to the presence of InDels in the two IS630-NC1 components of it.

As the first primary discovery of the preusent study, the first Tn630 member — Tn630-NC1 (Fig. 1A) exhibits significant distinctions from the well-known composite transposon Tn5 (GenBank: U00004), particularly regarding their gene structures. Discovered from *Escherichia coli*, Tn5 consists of two 1533-bp IS50 components (named IS50L and IS50R) and three genes within a 2752-bp region between them (Fig. 1B), whereas Tn630-NC1 consists of two 888-bp IS630-NC1 components (named IS630L and IS630R) and four genes in a 4685-bp region (Fig. 1A). The lengths of Tn5 and Tn630-NC1 are 5,818 bp and 6,461 bp, respectively. IS50L consists of a 19-bp 5' outside end of inverted repeat (OE), a 60-bp 5' UTR, a 1365-bp ORF of transposase (455 aa), a 70-bp 3' UTR, and a 19-bp 3' inside end of inverted repeat (IE), while IS50R consists of a 19-bp 5' OE, a 60-bp 5' UTR, a 1443-bp ORF of transposase (480 aa), and a 19 bp 3' IE sharing a 8-bp overlap with the ORF. Notably, the ORF in IS50L contains a G to T substitution, resulting in a PTC (Supplementary file 1). The three genes between IS50L and IS50R are kanamycin/neomycin, bleomycin and streptomycin resistance genes (denoted as kan, bleo and str), encoding proteins with lengths of 264, 129, and 267 aa, respectively. IS630R consists of a 21-bp 5' IE, a 12-bp 5' UTR, a 828-bp recovered ORF (275 aa), and a 21-bp 3' OE, while IS630L consists of a 21-bp 5' IE, a 12-bp 5' UTR, a 828-bp recovered ORF (275 aa), and a 21-bp 3' OE. The ORFs in IS630L and IS630R can be recovered by inserting "AA" to obtain the aa sequences of transposases (Supplementary file 1). The four genes between IS630L and IS630R encode four putative proteins with lengths of 791, 279, 91 and 90 aa, which were named as P1-4 (Fig. 1A), respectively. P1 and P2 have a 1-bp overlap in their ORFs, while P3 and P4 have a 20-bp overlap in their ORFs. P1 to P4 have been identified as belonging to VapE (Virulence-associated protein E, Genbank WP_052752335.1), the AntA/AntB antirepressor family (Genbank: WP_052752336.1), BrnT (the type II toxin-antitoxin system, Genbank: WP_046851183.1), and the BrnA antitoxin family (Genbank: WP_046851184.1), respectively. The significant distinctions in the structures exhibit that the genes encoding Tn630-NC1 and Tn5 transposases are transcribed in two different directions, one from the transposon bodies to the outsides, and the other from the outsides to the transposon bodies (Fig. 1A,B). The regulation of their transcription dependents on different surrounding sequences. Particularly, the transcription initiation of Tn630-NC1 transposases does not depend on the surrounding sequences outside Tn630-NC1, suggesting a higher degree of autonomy.

Discovery of the closest homolog of IS630 from viruses

Using the IS630 transposase for homology search (Methods and Materials), a special Tc1 member (BK032097: 1-1159) from phages (*Caudoviricetes sp*) was identified in the NCBI NT database. As the second primary discovery, this member was named as Tc1-C#1, using the name format of Tc1-XYn (**Described above**). Both Tc1-C#1 and IS630 (GenBank: X05955) consist of a 20-bp 5' TIR, a 84-bp 5' UTR, a 1,032-bp ORF (343 aa), a 3-bp 3' UTR, and a 20-bp 3' TIR (Supplementary file 1). Although both Tc1-C#1 and IS630 are not flanked by "TATA", they contain highly similar 20-bp TIRs at their 5' and 3' ends for identification of their boundaries (Fig. 1C). The 5' and 3' TIRs of Tc1-C#1 "CTAAATAGCTGCGC**CA**AATA" and "TATTA**G**GCGCAGCTATTTAG" contain two and one single nucleotide polymorphisms (SNPs), compared to "CTAAATAGCTGC

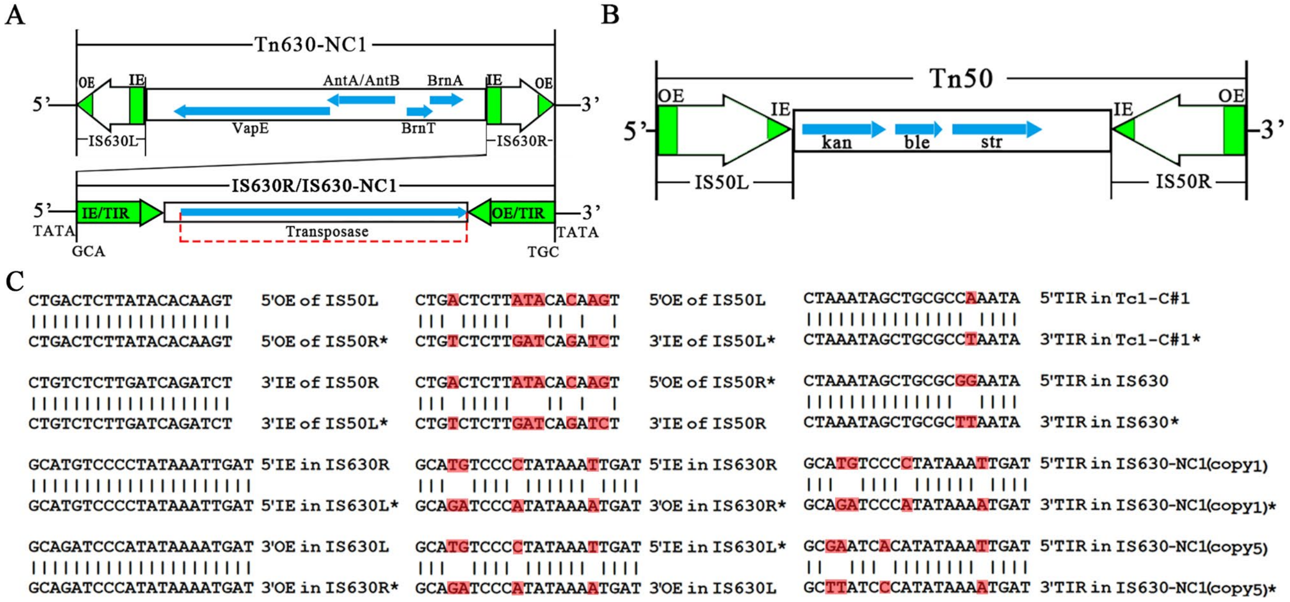
A
Tn630-NC1
OE
IE
AntA/AntB
BrnA
IE
OE
5'
3'
VapE
BrnT
IS630L
IS630R
IS630R/IS630-NC1
5'
IE/TIR
OE/TIR
3'
TATA
Transposase
TATA
GCA
TGC
B
Tn50
OE
IE
IE
OE
5'
3'
kan
ble
str
IS50L
IS50R
C
CTGACTCTTATACACAAGT 5'OE of IS50L
CTGACTCTTATACACAAGT 5'OE of IS50R*
CTGTCTCTTGATCAGATCT 3'IE of IS50R
CTGTCTCTTGATCAGATCT 3'IE of IS50L*
GCATGTCCCCTATAAATTGAT 5'IE in IS630R
GCATGTCCCCTATAAATTGAT 5'IE in IS630L*
GCAGATCCCATATAAAATGAT 3'OE in IS630L
GCAGATCCCATATAAAATGAT 3'OE in IS630R*
CTGACTCTTATACACAAGT 5'OE of IS50L
CTGTCTCTTGATCAGATCT 3'IE of IS50L*
CTGACTCTTATACACAAGT 5'OE of IS50R*
CTGTCTCTTGATCAGATCT 3'IE of IS50R
GCATGTCCCCTATAAATTGAT 5'IE in IS630R
GCAGATCCCATATAAAATGAT 3'OE in IS630R*
GCATGTCCCCTATAAATTGAT 5'IE in IS630L*
GCAGATCCCATATAAAATGAT 3'OE in IS630L
CTAAATAGCTGCGCCAAATA 5'TIR in Tc1-C#1
CTAAATAGCTGCGCCTAATA 3'TIR in Tc1-C#1*
CTAAATAGCTGCGCGGAATA 5'TIR in IS630
CTAAATAGCTGCGCTTAATA 3'TIR in IS630*
GCATGTCCCCTATAAATTGAT 5'TIR in IS630-NC1(copy1)
GCAGATCCCATATAAAATGAT 3'TIR in IS630-NC1(copy1)*
GCGAATCACATATAAATTGAT 5'TIR in IS630-NC1(copy5)
GCTTATCCCATATAAAATGAT 3'TIR in IS630-NC1(copy5)*

Fig. 1. Discovery of the first Tn630 member. (**A**) The first Tn630 member—Tn630-NC1 (NC represents *Nitrosomonas communis*) was discovered in the genome of *N. communis* strain Nm2 (GenBank: CP011451.1) from the NCBI NT database. The four genes between IS630L and IS630R include *VapE*, *AntA/AntB*, *BrnT* and *BrnA* (in blue color). (**B**) A well known composite transponson Tn5 (GenBank: U00004) had been discovered in *Escherichia coli* in a previous study. The three genes between IS50L and IS50R include *kan*, *ble* and *str* (in blue color). TIR: terminal inverted repeat (in green color); OE/IE: the outside/inside end of inverted repeat (in green color). (**C**) All the sequences start from 5 ' to 3'. * Read along the strand antisense to the reference of the transposon.

GC**GG**AATA" and "TATTA**A**GCGCAGCTATTTAG" of IS630, respectively. A previous study reported that another transposon — IS607 does not possess TIRs and does not necessarily generate target site duplications (TSDs) upon transposition[12]. Like IS607 and many other IS elements, IS630, IS50, Tn5, and Tc1-C#1 donot contain "TA" at their 5' and 3' ends. In contrast, IS630-NC1 and Tn630-NC1 do contain "TATA" at their 5' and 3' ends, suggesting the presence of TSDs. Therefore, it is still unclear whether the generation of TSDs is an essential prerequisite for transposition of IS630 or Tn630 members.

Subsequently, an extraordinarily close relationship between Tc1-C#1 and IS630 was delineated by: (1) a nucleotide (nt) identity of 88.09% (1021/1159); (2) highly similar 20-bp TIRs at their 5' and 3' ends; and (3) the presence of their transposases with an aa identity of 92.71% (318/343) and a positive substitution percentage of 95.92% (329/343); and (4) particularly, their DDE domains (the aa sequence of a DDE domain is specified to include residues from the first D to the third E) with an aa identity of 94.87% (111/117) and a positive-substitution percentage of 95.73% (112/117). According to the International Committee on Taxonomy of Viruses (ICTV) classification framework, *Caudoviricetes* is categorized as a class including many phage families[13]. As the host of IS630—*S. sonnei*, *Caudoviricetes* bacteriophages reside in human gut[14]. When phages infect bacteria, they may introduce their genetic material into the bacterial genomes during the infection process. As *S. sonnei* often infect human gut, causing severe diarrheal disease — bacterial dysentery[15], *Caudoviricetes* has potential to infect *S. sonnei.* Both the extraordinarily close relationship between Tc1-C#1 and IS630 and the relationship between their hosts strongly supported the occurrence of HT events involving Tc1-C#1, IS630, and their homologs between *Caudoviricetes* and *S. sonnei*. This suggested a dynamic exchange of genetic elements between gut bacteria and phages.

Using the IS630-AB1 transposase for homology search (Methods and materials), 12 new Tc1 members from viruses (Table 1) were identified in the NCBI NT database, including six with intact ORFs of transposases, two with partial ORFs, and four with ORFs containing InDels that lead to PTCs. Two of these InDels occurred within STRs, confirming our assumption that PTCs in the transposase ORFs of ITm members are likely predominantly caused by InDels within STRs, particularly within polyA regions. Through manual curation, one was excluded from the 12 members, based on the identification of its host (GenBank: KY052857) as the bacterium *Moraxella osloensis*, instead of an uncultured virus, as originally recorded in the NCBI GenBank database. Comparision of the remaining 11 Tc1 members with the viral homologs of IS630 led to a new result: the identities of the DDE domains between IS630 and its closely related homologs from viruses were substantially higher than those between other IS630 members (e.g., IS630-AB1) and their closely related homologs from viruses. For example, the highest identity of the DDE domains between IS630 and Tc1-C#1 reached 94.87%, followed by 80.34% between IS630 and its homologs from *Enterobacteria* phages and Stx2a-converting phages. In contrast, the identities of the DDE domains between IS630-AB1 and all its closely related homologs from *bamfordvirae* or *heunggongvirae* were below 35% (Table 1). Subsequent analysis of the closely related homologs of IS630 and IS630-AB1 revealed that all the viral hosts of them are phages except bodo saltans virus (BsV).

The Tc1 member from BsV (Table 1), named Tc1-BS1, has at least three copies in its genome (Genbank: MF782455). Discovered in 2017, BsV has been considered part of the most abundant group present in the sea. As the largest virus isolated so far, BsV has a genome with the size of about 1.39 Mb. As for the origin of BsV, one hypothesis is that BsV originated from bacteria that lost the ability to reproduce on their own and

ACC	Species	Kingdom	DNA type	Length	ORF size	AA number	ORF status	DDxE	Identity %
BK056286	*Bacteriophage sp.*	Bamfordvirae	Linear	39,150	849#	282	314DelA	DD34E	34.48
KY684085	*Indivirus ILV1*	Bamfordvirae	Linear	267,262	942	313	Intact	DD34E	23.53
KY684086	*Indivirus ILV1*	Bamfordvirae	Linear	147,851	939	312	Intact	DD34E	27.12
MF782455	*Bodo saltans virus*	Bamfordvirae	Linear	1,385,869	918	305	Intact	DD34E	28.81
MK072042	*Dasosvirus sp.*	Bamfordvirae	Linear	42,802	855	284	Intact	DD34E	28.45
KY322437	*Tetraselmis virus*	Bamfordvirae	Circular	668,031	918#	305	10InsA	DD34E	26.27
OL828820	*Chlorella virus*	Bamfordvirae	Circular	407,612	984	327	Intact	DD34E	29.41
BK031860	*Siphoviridae sp.*	Heunggongvirae	Linear	95,600	918#	305	517InsT	DD34E	28.93
KF114876	*Leptospira phage*	Heunggongvirae	Linear	86,537	1026	341	Intact	DD36E	26.27
BK038422	*Herelleviridae sp.*	Heunggongvirae	Linear	12,283	–	–	Partial	DD34E	48.28
KY695240	*Wolbachia phage*	Heunggongvirae	Linear	9054	–	–	Partial	DD31E	29.75
KY052857	*Uncultured virus**	Environmental sample	Linear	50,534	837#	278	294InsA	DD34E	62.21

Table 1. IS630/Tc1 members from viruses. A total of 12 IS630/Tc1 members were identified from viruses, using the IS630-AB1 transposase for homology search. A total of 10 columns include the accession number (**the 1st column**) of each virus nt sequence in the GenBank database, species (**the 2nd column**), kingdom (**the 3rd column**), DNA type (**the 4th column**), nt sequence length of the virus (**the 5th column**), the size of transposase ORF (**the 6th column**), the length of the aa sequence translated from the transposase ORF (**the 7th column**), the transposase ORF status (**the 8th column**), the DDxE motif (**the 9th column**), and the identity between the amino-acid sequence of the DDE domain in the IS630-AB1 transposase with that in each of the 12 IS630/Tc1 members (**the 10th column**). *The host was identified as the bacterium *Moraxella osloensis*, instead of an uncultured virus, as originally recorded. #ORFs were recovered by inserting or deleting nucleotides for translation.

became viruses, while the other hypothesis is that BsV originated from normal-size viruses that acquired a large amount of genes from other organisms during their evolution. Using the DDE domain for homology search in bacteria and other viruses, the closest homologs of Tc1-BS1 were identified from *Flavobacteriaceae bacterium* and *Indivirus ILV*, respectively. The DDE domain from *Flavobacteriaceae bacterium* had an aa identity of 38.46% (45/117) and a positive substitution percentage of 63.25% (74/117) with that of Tc1-BS1, while the DDE domain from *Indivirus ILV* had an aa identity of 38.79% (45/116) and a positive substitution percentage of 58.62% (68/116). Based on informaion of Tc1 and IS630 members up to the present date, the highest identity of the DDE domains between BsV and bacteria is comparable to that between BsV and other viruses (38.46% vs. 38.79%). Consequently, the origin of BsV, whether from bacteria or normal-size viruses, remains undetermined.

A model to elucidate the formation of composite transposons

When simple transposons are coupled into composite transposons (e.g., Tn630-NC1), TIRs that flank the simple transposons turn into OEs or IEs, depending on their specific positions (Fig. 1A,B). Our survey showed that: (1) for an intact simple transposon (e.g., IS50 or IS630), its 5' and 3' TIRs (denoted as TIR/TIR) tend to be identical, a condition also known as perfectly matched; and (2) for an intact composite transposon (e.g., Tn5 or Tn630-NC1), its OE pairs (denoted as OE/OE) and IE pairs (denoted as IE/IE) tend to be perfectly matched, respectively, while the paires composed of both OE and IE (denoted as OE/IE) tend to contain more mismatchs. For example, OE/OEs and IE/IEs within Tn630-NC1 and Tn5 are perfectly matched, respectively, whereas TIR/TIRs and OE/IEs within Tc1-C#1, IS630, IS630-NC1, and IS50 contain 1, 2, 4, and 7 mismatchs, respectively (Fig. 1C). Mismatchs in TIR/TIRs and OE/IEs within these simple transposons could have resulted from mutations, which provided a clue to understand the formation of composite transposons. According to previous knowledge, during the transposition of a composite or simple transposon, its transposase (Tnp) binds to its OEs or TIRs, forming Tnp-OE or Tnp-TIR complexes. Subsequently, the two complexes join together, and the C-terminus of Tnp interact and dimerize to form a synaptic complex with the ability to cleave DNA. By incorporating our discoveries into existing knowledge, we proposed a model to elucidate the formation of composite transposons in bacteria. This model was named as the "asymmetric TIRs" model. Mutations can cause one TIR (defined as the degenerated TIR) of a simple transposon a reduction or even loss of its binding abilities, while the other TIR (defined as the intact TIR) retains comparatively stronger binding abilities. When two copies of a simple transposon transpose into a small genomic region, they have potential to transpose together as a composite transposon. This requires that the intact TIRs of these two copies turn into OEs of the potential composite transposon, while the degenerated TIRs turn into IEs. The proximity of the intact OE/OE increases the likelihood that the inside composite transposon will transpose as a single unit, which is more likely than the separate transposition of the two copies. According to this model, the perfect matched OE/OEs and IE/IEs of Tn630-NC1 and Tn5 suggested their comparative intactness, while the mismatches in their OE/IEs indicated asymmetric TIRs (Fig. 1C). However, the mechanisms responsible for inactivation of these Tc1 or IS630 members are still not well understood. In particular, it remains unclear whether the inactivation process initiates with degeneration of TIRs or loss of transposase functionality.

Organization of IS630/Tc1 members into groups across biological kingdoms

Our survey, using the NCBI NT and NR databases, revealed that Tc1 and IS630 members have a broad host range, spanning across all six biological kingdoms (bacteria, fungi, plantae, animalia, archaea and protista) and viruses. Huge amounts of latent Tc1 and IS630 members in public databases have not been identified and their research values were neglected. A notable example is Tc1-S#1 (Genbank: CAE7467881) from *Symbiodinium sp*. (Table 2), suggesting the presence of a significant number of Tc1 members in protista. To comprehend the evolution of the Tc1 and IS630 families in a broader context, the closely related members across kingdoms can be organized into an IS630/Tc1 group. However, the sheer number of members within such a group is considerable and continues to grow with more IS630/Tc1 members identified. To effectively represent this diverse group, our strategy is to select seven members from the six kingdoms and viruses as representatives, respectively. In the present study, we proposed that any Tc1 or IS630 member can be classified into either an existing or new IS630/Tc1 group, by a unified procedure: (1) if a new Tc1 or IS630 member encodes a DDE domain that can be entirely (100%) covered by any member of existing groups, it will be assigned to the existing group as a closely related homolog, and if not, it will be assigned as the first member of a new group; (2) using the first member for homology search, an IS630 homolog from bacteria need to be selected as the IS630 representative of this group, based on the highest identity between its DDE domain and that of the first member; (3) using the IS630 representative for homology search, six other representatives from fungi, plantae, animalia, archaea, protista and viruses can be determined, if their DDE domains are 100% covered by the IS630 representative; (4) if more than one member from a kingdom are qualified to be a representative for the kingdom, only the member with the highest identity between its DDE domain and that of the IS630 representative will be selected; and (5) a group including seven representatives from six kingdoms and viruses is defined as a complete IS630/Tc1 group.

Using this procedure, the first IS630/Tc1 group (Table 2) was constructed and represented by seven representatives, including IS630-AB1 from *Acinetobacter baumannii* (GenBank: CP044356), Tc1-BS1 from Bodo saltans virus (GenBank: MF782455), Tc1-RS1 from *Rhizoctonia solani* (GenBank: KEP50069), Tc1-CC1 from *Cinara cedri* (GenBank: VVC27352), Tc1-QS1 from *Quercus suber* (GenBank: POF13514), Tc1-N#1 from *Nitrososphaera sp.* (GenBank: MDE1816798) and Tc1-S#1 from *Symbiodinium sp.* (GenBank: CAE7467881). Although IS630-NC1 had been assigned to this group, it was not selected as the representative, as IS630-AB1 had already been selected as the representative for bacteria. Tc1-MP1 and Tc1-OP1 had been identified by homology search using IS630-AB1[3], however, they were not assigned to this group. Instead, they had been assigned to another IS630/Tc1 group marked by IS630-PM1 (GenBank: MCL8582695) from *Proteus mirabilis.* A total of three IS630/Tc1 groups (Table 2) were intensively analyzed in the present study and they are: (1) group 0, marked

Member	Group	Species	Kingdom	ACC	DDxE	Identity%
IS630	0	*Shigella sonnei*	Bacteria	X05955	DD35E	100
Tc1-C#1		*Caudoviricetes sp.*	Virus	BK032097	DD35E	94.87
Tc1-MS1		*Mucor saturninus*	Fungi	KAG2190898	DD34E	26.67
Tc1-CT1		*Cyprideis torosa*	Animalia	OB689799*	DD37E	52.5
–		–	Plantae	–	–	–
Tc1-M#1		*Methanothrix sp.*	Archaea	MDD1748280	DD36E	37.82
–		–	Protista	–	–	–
IS630-BT1	1	*Bacillus thuringiensis*	Bacteria	OUB07123	DD34E	100
Tc1-KB1		*Kitale bracovirus*	Virus	EF710630*	DD34E	28.24
Tc1-EB1		*Enteropsectra breve*	Fungi	KAI5151414	DD34E	51,18
Tc1		*Caenorhabditis elegans*	Animalia	X01005	DD34E	52.76
Tc1-RL1		*Rhytidiadelphus loreus*	Plantae	OX344763*	DD34E	48.06
Tc-##1		Unknown	Archaea	RYH00493	DD34E	40.74
–		–	Protista	–	–	–
IS630-AB1	2	*Acinetobacter baumannii*	Bacteria	CP044356*	DD34E	100
Tc1-BS1		*Bodo saltans virus*	Virus	MF782455*	DD34E	28.81
Tc1-RS1		*Rhizoctonia solani*	Fungi	KEP50069	DD34E	37.19
Tc1-CC1		*Cinara cedri*	Animalia	VVC27352	DD34E	37.07
Tc1-QS1		*Quercus suber*	Plantae	POF13514	DD34E	35.83
Tc1-N#1		*Nitrososphaera sp.*	Archaea	MDE1816798	DD34E	36.21
Tc1-S#1		*Symbiodinium sp.*	Protista	CAE7467881	DD34E	33.62

Table 2. Members of three IS630/Tc1 groups. Each of 18 members in three IS630/Tc1 groups (named 0, 1, and 2) was described by its name (**the 1st column**), group number (**the 2nd column**), species (**the 3rd column**), kingdom (**the 4th column**), sequence accession number (**the 5th column**), the DDxE motif (**the 6th column**), and the identity between the amino-acid sequence of the DDE domain of each member and that of the bacterial member in the group (**the 7th column**). The accession numbers marked by * represent nt sequences and others represent aa sequences.

by IS630 (GenBank: X05955); (2) group 1, marked by Tc1 (GenBank: X01005); and (3) group 2, marked by IS630-AB1, which is the first IS630/Tc1 group constructed using our procedure. Additionally, group 3, marked by IS630-PM1, was intensively analyzed in our previous study[3]. Group 0, 1, 2 and 3 were also named as the Tc1, IS630, IS630-AB1, and IS630-PM1 groups. Theoretically, all IS630/Tc1 groups classified using our procedure are complete groups, however, it's noteworthy that most of these groups do not include members from hosts covering the six kingdoms and viruses. This limitation arises from insufficient data available in the NCBI NT or NR databases for a comprehensive representation across diverse organisms. Therefore, among the four groups (0, 1, 2, and 3), only group 2 is a complete group up to the present date and is also the first complete group under the ITm superfamily.

Phylogenetic analysis using the DDE domains (Methods and Materials) resulted in unexpected results (Fig. 2). The 18 representatives in group 0, 1 and 2 were not able to be clustered into two clades, corresponding to the Tc1 and IS630 families. Instead, they were clustered into three clades, corresponding to the three groups (0, 1 and 2). The similar results were also obtained using additional groups (data not shown), indicating that, as a whole, IS630 members cannot be distinguished from Tc1 members based on the analyzed features. Historically, the IS630 family was defined, based on several reasons, one of which was the difference in the DDxE motif of IS630 (DD35E) compared to Tc1 (DD34E). However, it was proposed that the DDxE motif is not a highly conserved feature for the classification of ITm members, a conclusion supported by previous studies and confirmed by the present study. The previous studies reported DD37E, DD37D, DD38E, DD39D[7], and DD40E[3], one after another, and the present study reported a large number of IS630/Tc1 members containing DDxE ($x > 34$), particularly four (IS630, Tc1-C#1, Tc1-CT1 and Tc1-M#1) of the five representatives in group 0 containing DD35E, DD35E, DD37E, and DD36E, respectively (Table 2). Based on these results, we concluded that IS630 should not be defined as a distinct family. Instead, it should be merged with Tc1 into a single family, named IS630/Tc1. We recommended that members of IS630/Tc1 from bacterial species and those from non-bacterial species can still be named using IS630-XYn and Tc1-XYn, respectively. As an additional finding, the 18 representatives in group 0, 1 and 2 were not able to be clustered into clades corresponding to six biological kingdoms and viruses. Multiple sequence alignment revealed that a significantly higher number of aa residues are conserved within group 0, 1, or 2 than those conserved within biological kingdoms (Fig. 3). The conserved aa residues in each group formed distinct blocks, which merits future investigation.

Further analysis of the three groups (0, 1, and 2), along with additional groups confirmed the aforementioned findings. Particularly, the identities of the DDE domains between IS630 and its viral representatives (*i.e.*, closely related homologs from viruses) are substantially higher than those between other IS630 members and their viral representatives; and in our studied IS630/Tc1 groups, the hosts of many viral representatives are phages.

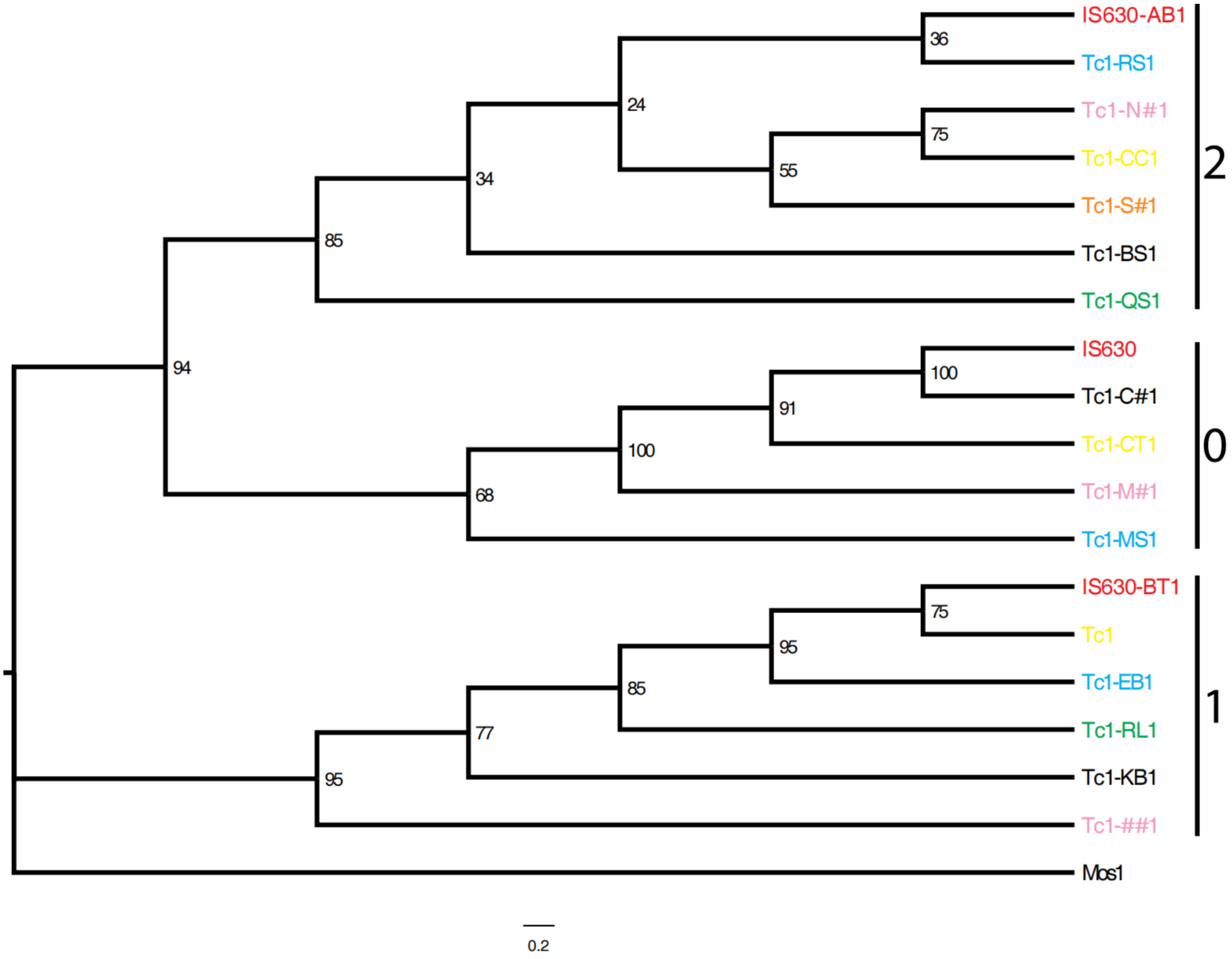

Fig. 2. Phylogenetic analysis of the 18 representatives in IS630/Tc1 group 0, 1 and 2. The maximum likelihood (ML) methods were used for phylogenetic analysis of the 18 representatives in IS630/Tc1 group 0, 1 and 2 (Table 2). The DDE domains (the aa sequence of a DDE domain is specified to include residues from the first D to the third E) were multiple aligned for the analysis. The species of the 18 representatives belongs to bacteria (red), fungi (blue), plantae (green), animalia (yellow), archaea (pink), protista (orange) and viruses (black). Numbers on the nodes represent bootstrap values expressed as percentages calculated using 10,000 bootstrap pseudoreplicates with a standard bootstrap of 1000 replicates. Mos1 (GenBank: X78906.1) was used to represent *Mariner* as outgroup.

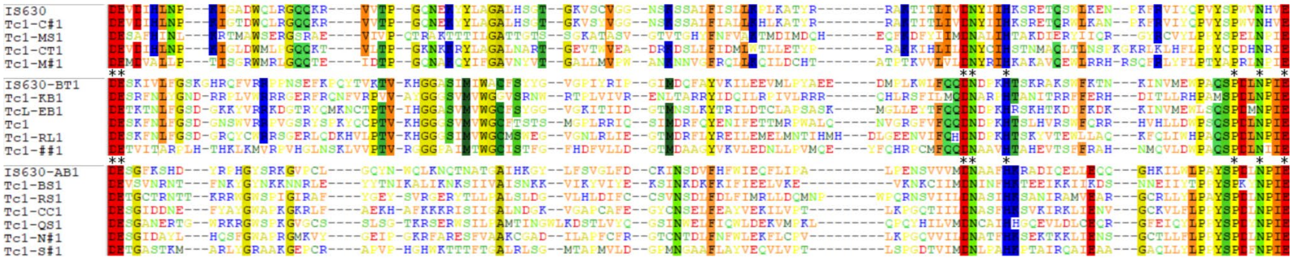

Fig. 3. Conserved blocks of amino-acid residues in IS630/Tc1 group 0, 1 and 2. Multiple alignment of the 18 representatives (Table 2) revealed that a significantly higher number of amino-acid (aa) residues are conserved in IS630/Tc1 group 0, 1, or 2 than those conserved in biological kingdoms. 20 amino-acid residues are represented using one-letter symbols in different colors, including: red for D and E; blue for K, R and H; green for N, Q, S, T, C and M with a gradient of color depth transitioning from lower to higher intensity; and yellow to orange for P, Y, W, G, A, F, L, V and I with a gradient of color depth transitioning from lower to higher intensity.

However, there are still some exceptions, due to insufficient data in the NCBI NT or NR databases. For example, Tc1-H#1 from *Herelleviridae sp.* should have been selected as the viral representative of the group 2, but it was excluded due to the presence of ambiguous residues (denoted as 'X') in the aa sequence of Tc1-H#1 transposase (GenBank: DAO92945). Thus, Tc1-BS1 from Bodo saltans virus (Table 2) was chosen instead. Unfortunately, the identity of the DDE domains between Tc1-BS1 and IS630-AB1 was only 28.81%, much lower than that between Tc1-H#1 and IS630-AB1, which was estimated between 48.28 and 56.9%. The above results will guide our future exploration of potentially active IS630/Tc1 members in group 0, to verify: (1) HT of Tc1-C#1 may still be occurring between *Caudoviricetes spp.* and *S. sonnei*; and (2) HT of other IS630/Tc1 members may still be occurring between gut bacteria and phages. Expanded exploration need be conducted to detect HT of other active transposons between gut bacteria and phages.

Conclusions

In the present study, our survey revealed that Tc1 and IS630 members have a broad host range, spanning across all six biological kingdoms (bacteria, fungi, plantae, animalia, archaea and protista) and viruses. In contrast, IS607[12] has been detected only in bacteria, fungi ((including *Ascomycetes* and *Basidiomycetes*), animalia, protista (e.g., *Amoebozoa* and *Stramenopiles*), and viruses. This deepened our understanding that the ITm superfamily represents the most widely distributed superfamily of DNA transposons. The primary discoveries include the first Tn630 member—Tn630-NC1 and the closest homolog of IS630 from viruses—Tc1-C#1. Further research on the first primary discovery led to the proposal that the formation of composite transposons may result from asymmetric TIRs. Further research on the second primary discovery revealed that IS630 and its homologs constitute a valuable resource for studying HGT between gut bacteria and phages, opening up new avenues for research in this field. Organization of Tc1 and IS630 members into groups across biological kingdoms facilitates data collection for future research, particularly on their HT between different kingdoms. By analyzing the IS630/Tc1 groups constructed by our established procedure, we reached several conclusions. A notable one is that IS630 should not be defined as a distinct family. Instead, it should be merged with Tc1 into a single family, named IS630/Tc1.

Although numerous cases of prokaryote-to-prokaryote and eukaryote-to-eukaryote HTs via transposons have been reported, only few have been documented between prokaryotes and eukaryotes[16]. This scarcity is likely due to transcriptional incompatibilities[17]. Theoretically, HT between phages and bacteria should be more common than that of prokaryote-to-prokaryote and eukaryote-to-eukaryote. due to the high frequency of phage infections. However, HT of IS630/Tc1 members between gut bacteria and phages may be more complex than expected, as it involves three possible routes (virus-to-virus, virus-to-bacteria and bacteria-to-bacteria). It is likely that phages act as vectors facilitating HT of IS630/Tc1 members between bacteria and bacteria, as previously proposed for other transposons[16]. Understanding the role and impact of transposons in genome evolution is a complex and fascinating area of research that can enhance our understanding of genetic diversity and adaptation. Unlike protein-coding genes, transposons generally do not provide any beneficial function to the genome, and their movement and proliferation usually have various negative effects[18]. As a result, there is a tendency for them to be inactivated by negative selection over evolutionary time. When homologs of a transposon are detected across different species, they typically share very low aa identities in their transposases. Therefore, the highest aa identity between IS630 and Tc1-C#1 transposases and the near-perfectly matched TIRs of Tc1-C#1 suggested that the discovery of Tc1-C#1 has significant implications. Our plans for future exploration of potentially active IS630/Tc1 members in group 0 can provide valuable insights into their dynamics and impacts in gut bacteria and phages, which could shed light on the ongoing evolutionary processes. If frequent HT of IS630/Tc1 members between bacteria and phages is confirmed, it could indeed suggest positive effects of transposons on genome evolution.

Methods and materials

All the nt or aa sequences were downloaded from the NCBI GenBank or RefSeq database for the analysis in local servers. Particularly, the nt sequences of Tc1, IS630, and *pogo* are specified to GenBank: X01005, GenBank: X05955, and GenBank: X59837, respectively, while those of Tc1-OP1, Tc1-MP1, and IS630-AB1 are located in the genomes of *O. polymorpha* DL-1 (CP080317: 544314–549972), *M. piriformis* (OW971871: 2795447 – 2793760), and *A. baumannii* CAM180-1 (CP044356: 173741–174626). Mos1 (GenBank: X78906.1) was used to represent *Mariner*. All other nt sequences of ITm members or the aa sequences of their transposases are provided in the Supplementary file 1. Six biological kingdoms (bacteria, fungi, plantae, animalia, archaea and protista) and viruses are specified to Bacteria (taxid: 2), Fungi (taxid: 4751), plants (taxid: 3193), Animalia (taxid: 33208), Archaea (taxid: 2157), algea (taxid: 3027), and Viruses (taxid: 10239) in the NCBI taxonomy database, respectively. The software BLAST v2.14.1 was used to search for homologs in a local NCBI NT and NR database with parameter setting (Word size = 3, Matrix = BLOSUM62, Gap existence = 11, Gap extension = 1). Multiple sequence alignment of the DDE domains was performed using Promal3D[19]. The maximum likelihood (ML) method was applied for phylogenetic analysis, using the software PhyloSuite[20] v1.2.2. Statistics and plotting were conducted using the software R v2.15.3 with the Bioconductor packages[21]. All other data processing was carried out using in-house Perl scripts.

Data availability

All data supporting the findings of this study are available within the paper and its Supplementary information.

Received: 28 August 2024; Accepted: 31 October 2024
Published online: 07 November 2024

References

1. Plasterk, R. H., Izsvák, Z. & Ivics, Z. Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet.* **15** (8), 326–332 (1999).
2. Emmons, S. W., Yesner, L. & Katzenberg, D. Evidence for a member in caenorhabditis elegans. *Cell* **32** (1), 55–65 (1983).
3. Chang, J. et al. The first discovery of Tc1 members in yeast. *Front. Microbiol.* **14**, 1–8 (2023).
4. Jacobson, J. W., Medhora, M. M. & Hartl, D. L. Molecular structure of a somatically unstable transposable element in Drosophila. *Proc. Natl. Acad. Sci.* **83** (22), 8684–8688 (1986).
5. Matsutani, S., Ohtsubo, H., Maeda, Y. & Ohtsubo, E. Isolation and characterization of IS elements repeated in the bacterial chromosome. *J. Mol. Biol.* **196** (3), 445–455 (1987).
6. Tudor, M., Lobocka, M., Goodell, M. & Pettitt, J. O'Hare, K. The pogo transposable element family of drosophila melanogaster. *Mol. Gen. Genet.* **232**, 126–134 (1992).
7. Shao, H. & Tu, Z. Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics* **159** (3), 1103–1115 (2001).
8. Gao, B. Prokaryotic and eukaryotic horizontal transfer of sailor (DD82E), a new superfamily of IS630-Tc1-Mariner DNA transposons. *Biology* **10** (10), 1005 (2021).
9. Chang, J. et al. Full-length genome of an ogataea polymorpha strain CBS4732 ura3Δ reveals large duplicated segments in subtelomeric regions. *Front. Microbiol.* **13**, 1–10 (2022).
10. Xu, X. et al. Using pan RNA-seq analysis to reveal the ubiquitous existence of 5' and 3' end small RNAs. *Front. Genet.* **10**, 1–11 (2019).
11. Dupeyron, M., Baril, T., Bass, C. & Hayward, A. Phylogenetic analysis of the Tc1/mariner superfamily reveals the unexplored diversity of pogo-like elements. *Mob. DNA* **11**, 21 (2020).
12. Kersulyte, D., Mukhopadhyay, A. K., Shirai, M., Nakazawa, T. & Berg, D. E. Functional organization and insertion specificity of IS607, a chimeric element of Helicobacter pylori. *J. Bacteriol.* **182**, 5300–5308 (2000).
13. Zhu, Y., Shang, J., Peng, C. & Sun, Y. Phage family classification under caudoviricetes: a review of current tools using the latest ICTV classification framework. *Front. Microbiol.* **16** (13), 1032186 (2022).
14. Benler, S., Yutin, N., Antipov, D., Raykov, M. & Koonin, E. V. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* **9**, 78 (2020).
15. Brunner, K., Samassa, F., Sansonetti, P. J. & Phalipon, A. Shigella-mediated immunosuppression in the human gut: subversion extends from innate to adaptive immune responses. *Hum. Vacc. Immunother.* **15** (6), 1317–1325 (2019).
16. Clément, G. & Cordaux, R. Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biol. Evol.* **5** (5), 822–832 (2013).
17. Gladyshev, E. A. & Arkhipova, I. R. A single-copy IS5-like transposon in the genome of a bdelloid rotifer. *Mol. Biol. Evol.* **26** (8), 1921–1929 (2009).
18. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
19. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36** (7), 2295–2300 (2008).
20. Zhang, D. et al. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* **20** (1), 348–355 (2020).
21. Gao, S., Ou, J. & Xiao, K. R language and Bioconductor in bioinformatics applications (Chinese Edition). *Tianjin: Tianjin Sci. Technol. Translation Publishing Ltd.* (2014).

Acknowledgements

We appreciate the special help from professors Wenjun Bu, Tao Zhang, Huaijun Xue, Bingjun He, Qiang Zhao, Wei Liu, and Zhen Ye from College of Life Sciences, Nankai University and Dr. Tung On Yau from Nottingham Trent University. This manuscript has been submitted as a preprint on Jan 23nd, 2024 at https://www.researchsquare.com/article/rs-3890757/v1.

Author contributions

SG conceived the project. SG and CR supervised the present study. GD, HY and JC performed the programming. YH, YG, MZ, and SY analyzed the data. WL prepared the figures, tables and supplementary files for submission. SG drafted the manuscript. SG and CR revised the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the Open Project of Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources (2024-KF-03) to Shan Gao and Kunlun Talent Program of Qinghai Province to Wenjing Li. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-78495-z.

Correspondence and requests for materials should be addressed to C.R. or S.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.