Sicat *et al. Mobile DNA* (2022) 13:12
https://doi.org/10.1186/s13100-022-00270-6

Mobile DNA

RESEARCH

Open Access

Characterization of transposable elements within the *Bemisia tabaci* species complex

Juan Paolo A. Sicat[1*], Paul Visendi[2], Steven O. Sewe[1], Sophie Bouvaine[1] and Susan E. Seal[1]

Abstract

Background: Whiteflies are agricultural pests that cause negative impacts globally to crop yields resulting at times in severe economic losses and food insecurity. The *Bemisia tabaci* whitefly species complex is the most damaging in terms of its broad crop host range and its ability to serve as vector for over 400 plant viruses. Genomes of whiteflies belonging to this species complex have provided valuable genomic data; however, transposable elements (TEs) within these genomes remain unexplored. This study provides the first accurate characterization of TE content within the *B. tabaci* species complex.

Results: This study identified that an average of 40.61% of the genomes of three whitefly species (MEAM1, MEDQ, and SSA-ECA) consists of TEs. The majority of the TEs identified were DNA transposons (22.85% average) while SINEs (0.14% average) were the least represented. This study also compared the TE content of the three whitefly genomes with three other hemipteran genomes and found significantly more DNA transposons and less LINEs in the whitefly genomes. A total of 63 TE superfamilies were identified to be present across the three whitefly species (39 DNA transposons, six LTR, 16 LINE, and two SINE). The sequences of the identified TEs were clustered which generated 5766 TE clusters. A total of 2707 clusters were identified as uniquely found within the whitefly genomes while none of the generated clusters were from both whitefly and non-whitefly TE sequences.

This study is the first to characterize TEs found within different *B. tabaci* species and has created a standardized annotation workflow that could be used to analyze future whitefly genomes.

Conclusion: This study is the first to characterize the landscape of TEs within the *B. tabaci* whitefly species complex. The characterization of these elements within the three whitefly genomes shows that TEs occupy significant portions of *B. tabaci* genomes, with DNA transposons representing the vast majority. This study also identified TE superfamilies and clusters of TE sequences of potential interest, providing essential information, and a framework for future TE studies within this species complex.

Keywords: Transposable elements, Whitefly, Bioinformatics, *Bemisia tabaci*, DNA transposons, TE annotation

Introduction

Whiteflies are agricultural pests that cause crop losses amounting to billions of dollars[1–3]. More than 1500 whitefly species have been identified and amongst them, the members of *Bemisia tabaci* whitefly species complex are the most damaging collectively in terms of their broad crop host range (e.g. beans, cassava, cotton, potato, tomato) and ability to serve as a vector for > 400 plant viruses [4–6].

Agricultural intensification and climate change have led to highly fecund populations of *B. tabaci* spreading across continents and globally through international trade of infested plants [1, 7, 8]. The severity of this pest species complex has for decades shaped several national and international collaborative projects [3], with a dramatic

*Correspondence: J.A.Sicat@greenwich.ac.uk
[1] Natural Resources Institute, University of Greenwich, Central Avenue, Gillingham, Chatham ME4 4TB, UK
Full list of author information is available at the end of the article

increase in genome and transcriptome resources in the past decade. These have assisted in the exploration of mechanisms that underly diversification within this pest species complex, such as differing host specificities and detoxification mechanisms, and plant virus interactions [9–13]. In the last few years, draft genome assemblies have been published (MEAM1, MED/Q, and SSA-ECA) alongside the annotation of genomic features that are associated with insecticide resistance, detoxification, and virus transmission [14–16]. Transposable elements (TEs) have, however, been neglected in all of these studies with no detailed characterization to date of TEs found in this whitefly species complex.

The identification of TEs is integral in the analysis of genome assemblies as TEs are abundant in eukaryotic genomes and can multiply, move, affect gene regulation, and expand the host's genome [17–21]. TEs are classified into two main classes based on their method of transposition: DNA transposons and Retrotransposons [22–25]. DNA transposons transpose with the aid of a DNA intermediate and can either be autonomous or non-autonomous [23, 26, 27]. Autonomous elements can transpose on their own while the non-autonomous elements require other TEs to facilitate their movement [27, 28]. The majority of DNA transposons utilize a "cut-and-paste" method of transposition; wherein the transposons are "cut" from their position and then "pasted" (inserted) into a new target site [18, 28, 29].

Retrotransposons are TEs that can transpose with the aid of an RNA intermediate [30–32]. While DNA transposons encode for a transposase, retrotransposons produce RNA transcripts, and they are transcribed from RNA to DNA with the aid of reverse transcriptase enzymes and the sequence is then integrated into new sites in the genome [30, 31]. Their mobilization in the genome does not require excision hence their movement has been dubbed as "copy-and-paste" [31, 33]. Retroelements can be further classified based on their structures into two orders: Long terminal repeat (LTR) retrotransposons and NonLTR retrotransposons [30–32].

TEs can be further classified into superfamilies and their presence in the different arthropod species varies greatly, currently ranging from as low as 2.6% in *Belgica antartica* to as high as 72.8% in *Sitophilus oryzae* [34, 35]. The functions of these elements are often unknown, but their presence in genomes has been associated with inducing various changes in their host organism. The majority of TE studies in insects have been in drosophilids and one of the most characterized TE is the P element [36, 37]. P elements were first discovered in *Drosophila melanogaster* and were shown to cause hybrid dysgenesis [38], which occurs when female strains of *D. melanogaster* that lack P elements mate with male strains with autonomous P elements [36, 39]. The resulting combination results in progeny with sterility disorders, an elevated mutation rate, and increases in chromosomal rearrangements and recombination [36, 39, 40]. Different types of TEs have different effects and the characterization of these elements in other insect species has underpinned an improvement in our understanding of the potential impacts of these elements.

The roles that TEs can play in gene regulation and expression have already been described [28, 41–46] and the abundance and types of TEs in the different whitefly genomes could have shaped the evolution of the species complex. TEs have also been associated with gene duplication wherein the insertion location of the TE affects the normal replication process [17, 47]. The exact mechanism of the alteration of the process depends on the type of TE and the extent of its effects vary accordingly [44, 47–49].

TEs represent a major proportion of *B. tabaci* genomes, accounting for approximately 40–44% of the published draft genomes of two *B. tabaci* species termed Middle East Asia Minor 1 (MEAM1) and Mediterranean Q (MED/Q) [14, 16]. The latest released *B. tabaci* draft genome of a SubSaharan African population (SSA-ECA), reported a slightly lower (38.5%) TE content but it was noted that the 513 Mb genome assembly was missing around a quarter of genome data [15]. Hence the repeat content of the SSA-ECA genome cannot be considered as accurate.

Aside from the proportion of TEs found within the *B. tabaci* genomes, little is known on the TEs found within the *B. tabaci* species complex. In addition, there are marked differences in reported estimates of TE orders between the above two complete *B. tabaci* draft genomes. Although all the studies reported around ~ 40% of the genomes to be comprised of TEs, the MEAM1 and SSA-ECA whitefly genomes were reported to have an abundance of DNA transposons [14, 15] particularly MITEs (miniature inverted-repeat transposable elements) while LTRs were reported [16] to be the most abundant in the MED/Q genome. Members of the *B. tabaci* species complex show very different biological and phenotypic properties and hence these contrasting results were considered potentially significant.

The studies that reported very different TE class proportions in the *B. tabaci* whitefly genomes [14–16] employed different TE annotation workflows. In both MEAM1 and SSA-ECA annotation [14, 15], a MITE-specific identification tool was included (MITE-Hunter), whereas LTR-specific identification tool (LTR-Finder) was incorporated in the MED/Q repeat annotation workflow [16]. Chen et al. [14, 15] created their species-specific repeat libraries using RepeatModeler (RECON and

RepeatScout) and included MITE-Hunter for the identification of MITEs. Xie et al. [16] used Piler-DF, and RepeatScout to create their repeat library and included LTR-FINDER to identify LTRs.

The three whitefly draft genome assemblies used different genome sequencing technologies and assembly methods. MEAM1 whitefly DNA was sequenced using Illumina HiSeq 2500 system, and Illumina paired end reads assembled by Platanus v1.2.1, with gaps subsequently filled using PacBio long reads and PBJelly [14]. The MED/Q genome assembly was also constructed from Illumina paired end reads, but assembled using SOAPdenovo [16] followed by using Bacterial Artificial Chromosome (BAC) libraries to improve assemblies. For the most recently released SSA-ECA draft genome assembly, paired end and mate pair libraries from HiSeq 2500 were used with Platanus [15]. Pilon was included to fill in gaps. The SSA-ECA publication [16] noted that although ~ 25% of the genome was missing, the majority of the gene space was considered to have been assembled correctly.

The use of different assembly methods and workflows hinders the accurate comparisons of TE classes previously reported across the three *B. tabaci* genomes. Reliable inferences based on the significant differences in TE compositions found across the published genomes of the *B. tabaci* species complex can therefore not be made. Furthermore, attempts were made to replicate the identification workflows reported in the published data and results were inconsistent with the published estimates using the same genome assemblies. To address the issue of the assemblies using different TE annotation workflows, this study developed a reproducible workflow for identifying and classifying TEs found within *B. tabaci* genomes. The application of the same workflow across all the published *B. tabaci* genomes provided a standardized TE annotation process and highlighted some misclassification and an overestimate of TE compositions in currently published *B. tabaci* genomes. This study provides the first accurate exploration of TE classes in the *B. tabaci* species complex.

Results

Identification of TEs using the repeatmasker repBase library

The three draft genomes (MEAM1, MED/Q, and SSA-ECA) for the *B. tabaci* cryptic species complex published to date were the focus of analyses. TEs within these genomes were initially identified using a RepBase library (version RepBase_RepeatMasker-edition20180826 library) through RepeatMasker. The results of the TE identification using the RepeatMasker RepBase library were significantly lower than reported in their respective publications (Table 1); MEAM1 (18.92% vs 43.82% published), MED/Q (17.28% vs 40.29% published), and SSA-ECA (13.41% vs 38.52% published).

The RepBase library was searched for *B. tabaci*-specific TEs and 282 different TE consensus sequences were identified. The result of the identification showed that only some of the identified TE consensus sequences were submitted to RepBase and with these submitted consensus TE sequences, only less than half of the published TEs were identified. Attempts to find the rest of the consensus sequences in publicly available repositories were unsuccessful.

The RepBase library was therefore tested for its ability to identify TEs in a *Drosophila melanogaster* genome (release 6 [50]) to identify if the anomalies for the hemipteran genomes tested in this study were due to user errors. The RepBase library was able to identify 17.44% TE genome proportion while published results show that < 20% of the genome was identified as TEs in different Drosophila studies [51–54]. The results of the identification were thus in line with what was reported to be found in the species, confirming that the library was being searched correctly.

Table 1 Repetitive elements identified in the three whitefly genomes

	MEAM1			MED/Q			SSA-ECA		
	Published	RepBase	Custom Library	Published	RepBase	Custom Library	Published	RepBase	Custom Library
DNA	29.25	18.07	25.28	15.66	16.48	23.42	25.94	12.92	19.86
Retroelements		0.86	2.6		0.61	2.65		0.42	1.72
LINE	0.96	0.61	1.25	3.18	0.57	0.96	0.44	0.38	0.94
SINE	0.16	0.04	0.17	0.96	0.04	0.18	0.16	0.04	0.08
LTR	0.49	0.21	1.19	18.5	0.19	1.51	0.08	0.07	0.7
Unknown	12.96	0	16.26	1.99	0	14.81	11.9	0	15.22
Total	43.82	18.92	44.14	40.29	17.28	40.88	38.52	13.41	36.8

Results of the identification of TEs reported by their respected studies, using the last publicly available RepBase library (RepBase RepeatMasker-edition20180826), and the custom-built repeat library built using the workflow described in the study

The results of the TE identification using the RepeatMasker RepBase library showed that the library could not be used for the characterization and comparison of the TEs found within the whitefly genomes. To resolve the issue, an annotation workflow was developed to standardize the identification of the TEs across the whitefly genomes. This had varied in the published research that utilized different TE identification tools; MEAM1, and SSA-ECA [14, 15] used a DNA transposons specific tool, while MED/Q [16] used a LTR specific identification tool. Standardization of the annotation workflow is required for an accurate comparison of TEs across the three genomes. A species-specific custom-built repeat library was created for each genome studied using the same tools to identify and classify TEs within each genome. The identification of the TEs in the workflow combines several methods in the identification of elements: structural-based and de novo; while the classification of the identified elements uses sequence similarity, structural, and machine learning (for details see Methodology section).

The performance of the annotation workflow developed was validated using a well characterized genome to determine its suitability for annotating TEs in less well characterized insect genomes. The *D. melanogaster* genome (release 6 [50]) was chosen for the validation as it is known to be one of the most accurate in terms of its TE annotation with several iterations of reference genome releases and information on TEs released alongside these [50, 55]. The annotation workflow developed was compared against the RepeatMasker RepBase library as the latter uses a database that contains the updates from several TE studies and libraries that includes the TE annotation from the *D. melanogaster* genome releases [24, 56].

A total of 17.44% genome proportion of interspersed repeats was found in the *D. melanogaster* using the RepeatMasker library compared to 16.88% genome proportion of interspersed repeats was found using the species-specific custom-built library (Table 2). Most of the repeats found were LTRs and a difference of 0.46% in this category was seen between the RepeatMasker and custom-built libraries. The SINE class of elements was the least common; the RepeatMasker library identified 81 bp of SINEs while the custom-built library found none (0 bp). For DNA transposons a difference of 0.58% was observed between the two libraries, while a difference of 0.42% was observed in the detection of LINEs. The difference of < 1% of the total of TEs identified in the *D. melanogaster* genome and less than < 1% in each of the orders support the capability of the workflow developed in identifying TEs found within a genome.

Table 2 RepeatMasker output of RepeatMasker library and the species-specific custom-built library for the Drosophila melanogaster genome (release 6 [50])

	RepBase (%)	Custom Library (%)
DNA	1.79	1.21
LINE	4.93	4.50
SINE	< 0.001	0.00
LTR	10.68	10.22
Unclassified	0.04	0.34
Total Interspersed Repeats	17.44	16.88

Comparison of the results of the identification of TEs using RepeatMasker RepBase library and the species-specific repeat library in the *D. melanogaster* genome. The custom-built repeat library was built using the workflow described in the study

TEs in arthropod genomes

The validated developed workflow was used to identify the TE content of each of the target genomes (Fig. 1), resulting in a custom-built species-specific library for each of the genomes studied. Aside from the three whitefly genomes (MEAM1, MED/Q, and SSA-ECA), three further hemipteran genomes were included as a general comparison, namely *Acyrthosiphon pisum*, *Diaphorina citri,* and *Myzus persicae*. The three whitefly genomes all had a higher TE content (an average of 40.61% genome proportion of TEs) compared to each of the three non-whitefly genomes (an average of 25.01% TE genome proportion). MEAM1 had the highest TE content across the six genomes at 44.14% while the *A. pisum* assembly had the highest TE content amongst the non-whitefly genomes at 34.54%. The SSA-ECA draft genome (known to be missing ~25% genome data) had the lowest TE content amongst the whitefly genomes at 36.80%, over 2% higher than the TE content in the *A. pisum* genome assembly. The *M. persicae genome assembly* had the lowest TE content across the six genomes at 17.52%.

The relationship between assembly sizes of the six genomes and their TE content was tested using Spearman's rank rho correlation (Fig. 2). TE proportion was found to be positively correlated with assembly size ($r = 0.93$, $p = 0.006$). The highest TE content (44.14%) across the six genomes was in the MEAM1 genome (615 Mbp) while the smallest genome, the *M. persicae* genome assembly (347 Mbp) had the lowest TE content at 17.52%. Amongst the whitefly genomes, SSA-ECA has the smallest assembly size (538.48 Mbp) and the lowest TE genome proportion (36.80%).

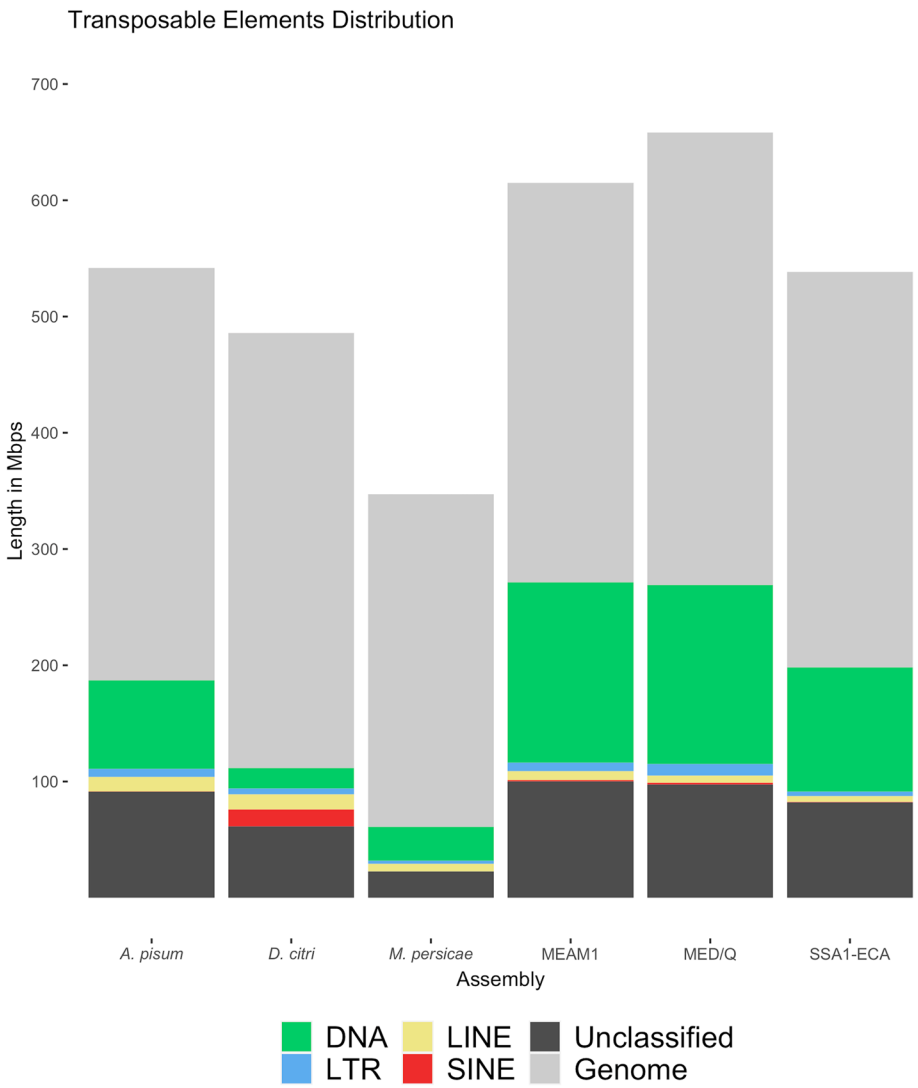

Fig. 1 Distribution of transposable elements in each respective genome. Stacked bar chart illustrating the length of each genome and the length occupied by the TEs in each genome. RepeatMasker with the species-specific repeat library was used to identify the TE content of each genome

Difference in the distribution of TE content between genomes

There was no statistically significant difference ($p = 0.09$) in genome assembly size between the whitefly genomes (average 603.92 Mbp) and the non-whitefly genomes (average 458.24 Mbp). This allowed us to compare the two groups without significantly biasing our results with the variations in genome assembly sizes. The distribution of TEs as a percentage of genome was compared across the six genomes. The majority of the classified elements within the whitefly genomes were DNA transposons at an average of 22.85% across the three genomes. MEAM1 had the highest distribution amongst the three whitefly genomes at 25.28% while SSA-ECA had the lowest at 19.86%. Retrotransposons were classified at a much lower average of 2.32% proportion in the whitefly genomes, with LTRs as the most abundant order identified across the three at an average of 1.13% followed by LINEs at an average of 1.05%.

For the three non-whitefly genomes, DNA transposons were the most abundant in the *A. pisum* (14.06%) and *M. persicae* (8.35%) genomes while retrotransposons were the most abundant class in the *D. citri* genome (6.68%). An average of 4.34% proportion was identified as retrotransposons within the non-whitefly genomes. LINEs were the most abundant retrotransposon order in the *A. pisum* genome(2.32%) and *M.*

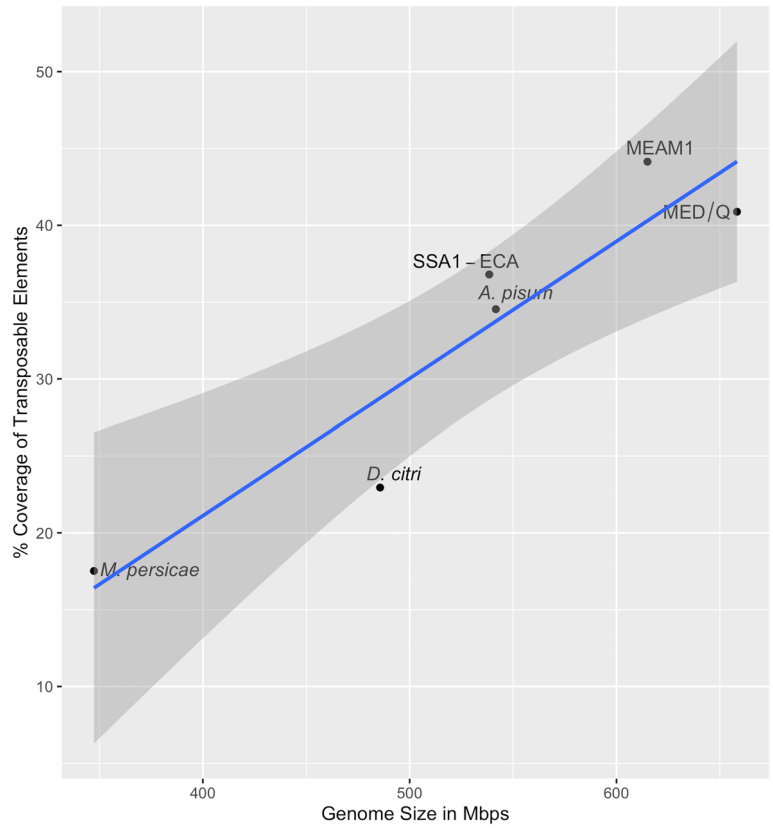

Fig. 2 Percent proportion of transposable elements and assembly size. Each genome was plotted in relation to their TE proportion and assembly size. TE proportion in the six genomes is positively correlated with the size of the genome assembly ($p = 0.006$). The grey shaded area represents the 95% confidence interval while the blue line is the regression line ($r = 0.0.93$)

persicae genome (1.86%) while SINEs were the most abundant in *D. citri* genome (3%).

Across the four orders of TEs, SINEs were the least identified at an average of 0.58% (0.14% for the whitefly genomes and 1.01% for the non-whitefly genomes). Amongst all the six genomes, the *D. citri* genome had the highest percentage of SINEs at 3% while this TE order was not detected in the *M. persicae* genome assembly.

The distribution of TEs between the genomes was explored further by comparing their distribution between the two groups of genomes to determine if there were any specific features that appeared to be specific to the whitefly genomes studied. The comparison of the distribution of the orders of the TEs between the whitefly and the non-whitefly genomes was performed using a two-sample t-test (DNA transposon, LTR, and LINE) and Wilcoxon rank-sum test (SINE) (Fig. 3). A standard t-test was used for orders that had the same variance (DNA transposons, LTRs, and LINEs) while a Wilcoxon rank-sum test for SINEs as the distribution for genome proportion in the two groups as they had a non-normal distribution. There is a significant difference between the mean TE content of DNA transposons ($p = 0.01$) and LINEs ($p = 0.008$) between the whitefly genomes and the non-whitefly genomes, while there was no significant difference found in LTRs ($p = 0.7856$) and SINEs ($p = 0.6625$). There are significantly more DNA transposons found in the whitefly genomes and significantly less LINEs compared to the three non-whitefly hemipteran genomes studied.

Unclassified elements are still found within the identified TEs. Across the six genomes, an average of 13.70% genome proportion remains unclassified (15.43% for the whitefly genomes and 11.98% for the non-whitefly genomes). The unknown consensus sequences from the whitefly species-specific TE libraries were searched against the NCBI non-redundant protein database and UniProtKB/Swiss-Prot Arthropoda protein sequences. The repeat sequences with hits were planned to be excluded from the final TE library; however, no matches were found.

Lastly, it should be noted that the relative proportions of the elements will be subject to change when the

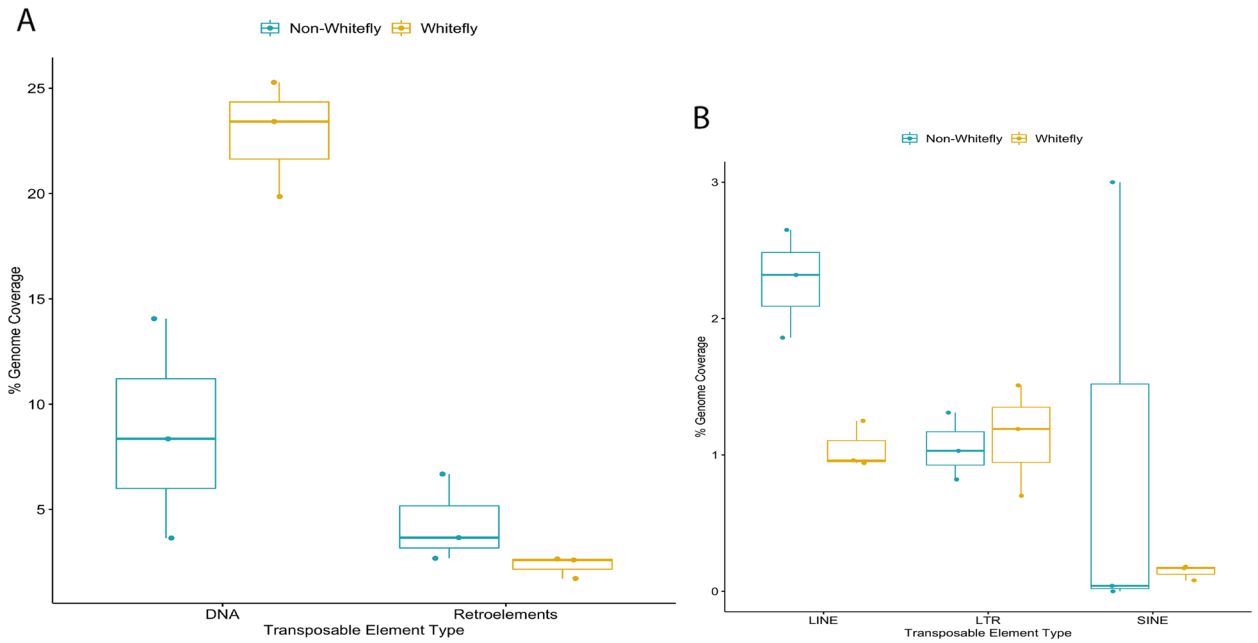

Fig. 3 Percent proportion of each order of transposable elements in the non-whitefly and whitefly group. Box plots comparing percent genome proportion of each order of TEs between the non-whitefly and the whitefly group. The box represents the interquartile range (25th to 75th percentile) values and the line in the middle of the box represents the middle quartile (50th percentile or median). The upper whisker represents the values 1.5 times larger than the 75th percentile and the lower whisker represents the values smaller than the 25th percentile. (A) The overview of the distribution of TEs classes between the non-whitefly and whitefly groups. The majority of the TEs identified were DNA transposons, most abundant in whitefly genomes. (B) The comparison of the distribution of retrotransposons between the non-whitefly and whitefly group. SINEs distribution varies significantly across the non-whitefly group; the *D. citri* genome assembly having the highest at 3% while none was detected in *M. persicae genome assembly*

Table 3 Repeat Superfamilies identified within the genomes

	DNA	LINE	LTR	SINE	Total
MEAM1	48	20	9	5	82
MED/Q	49	20	6	4	79
SSA-ECA	44	18	9	4	75
A. pisum	43	18	5	1	67
D. citri	30	23	6	7	66
M. periscae	36	18	7	0	61

The table presents a summary of the number of superfamilies found in each class of TEs in each of the genomes. DNA represent DNA transposons, *LINE* Long interspersed nuclear elements, *SINE* Short interspersed nuclear elements, *LTR* Long terminal repeats

unclassified elements become classified; nevertheless, the very high proportion of identified DNA transposons in the whitefly genomes means that this class will remain the largest order of elements identified within all three whitefly genomes analyzed (Supplementary Table 2).

TE superfamilies across the genomes

Each TE from the different orders can be further classified into superfamilies on the basis of their monophyletic origin and homology of motifs [27, 56, 57]. Superfamilies were identified in each genome (Table 3). A total of 98 TE superfamilies were identified in the whitefly genomes and 89 for the non-whitefly genomes. A total of 69 TE superfamilies were identified to be present across the genomes in the two groups (39 DNA transposon, eight LTR, 19 LINE, and three SINE). Most of the superfamilies identified were classified as DNA transposons with a total of 66 different superfamilies of which 19 were unique to whitefly genomes while eight were unique to non-whitefly genomes. SINE superfamilies were the least identified with 11 superfamilies of which four are unique to whitefly genomes and another four unique to the non-whitefly genomes. LINE superfamilies were the most identified retrotransposons with 29 unique superfamilies of which three are unique to whitefly genomes while seven are unique to the non-whitefly genomes.

MEAM1 showed the greatest number of superfamilies identified at 82 while the *M. persicae* genome had the lowest at 61 superfamilies. In all genomes, DNA transposon superfamilies were the most identified with an average of 47 in the whitefly genomes and 36 in the non-whitefly genomes. MED/Q and MEAM1 had the greatest number of DNA transposon superfamilies at 49 and 48 respectively, while the *D. citri* genome had the least

at 30 superfamilies. SINE superfamilies were the least identified at an average of four superfamilies. The *D. citri* genome had the greatest number of SINE superfamilies identified with seven while SINEs were not identified at all in *M. persicae* genome assembly.

Further analysis of the superfamilies found across the genome assemblies was performed by clustering the TE consensus from the six species-specific libraries. The clustering was based on the length of the TEs and 80% sequence similarity. A total of 5766 clusters were created; 1131 clusters from the non-whitefly TE consensus sequences, 2707 clusters from whitefly TE consensus sequences, and 1928 clusters created from TE consensus sequences found in the same genome assembly (Supplementary Table 3). The 1928 clusters from TE consensus sequences found in the same genome assembly were expected. These clusters were created from the same superfamilies found within the species-specific library of one genome (i.e. Gypsy element from MEAM1 identified as 80% similar to another MEAM1 Gypsy element). These expected overlaps are disregarded as clustering was already performed during the creation of the species-specific TE libraries (see Methodology). Although similar repeat superfamilies were identified across the genome assemblies based on their classification order, none of the TE consensus sequences from each group (whitefly vs. non-whitefly) was identified as shared based on their sequence similarity and length.

Breakdown of the 2707 clusters from the whitefly TEs (Table 4) reveals that MEAM1 and MED/Q genomes shared the greatest number of clusters at 987 while MED/Q and SSA-ECA shared the least at 441. A total of 733 clusters were identified as shared across the three *B. tabaci* genomes. There were 216 known TE clusters identified as DNA transposons of which 37 clusters were from Helitrons, 31 clusters were from different hAT families, and 31 were from different TcMar families. A total of 174 clusters are identified as LTRs and three superfamilies account for most of these clusters: 61 Copia clusters, 56 Gypsy clusters, and 54 Pao clusters. A total of 120 clusters are identified as LINEs and three superfamilies account for more than half of the clusters: 25 Jockey clusters, 19 L2 clusters, and 17 R1 clusters. Clusters classified as SINEs were identified the least, with only 2 SINEs clusters from the 733 clusters.

Lastly, a significant number (1004) of clusters from the whitefly TE clusters were from unclassified TE consensus sequences. Unclassified TE consensus sequences from MEAM1 and MED/Q created the greatest number of clusters at 318 clusters, followed by TEs from MEAM and SSA-ECA at 283 clusters and a total of 220 clusters were created from the three whitefly TE consensus sequences.

Repeat landscapes

With the help of a script included in RepeatMasker, several repeat landscapes were produced. These repeat landscapes show the sequence divergence measured by Kimura distance within each genome. The graphs below present the distribution of genome coverage of copies of each type of transposable element (DNA transposon, LTR, LINE, SINEs, Unknown, and others) and its divergence from their consensus sequence. A copy's divergence can infer its activity and age of insertion. A low divergence score implies a more recent transposable element activity while more divergent scores represent copies with older transposition events. Peaks of activity can also be observed in these graphs, and they represent transposition bursts in the evolutionary history of the specific transposable elements [58].

Figure 4 displays the repeat landscapes of the *B. tabaci* genome assemblies. SSA-ECA shows that the genome has one prolonged increase of TE activity around 5 to 10 Kimura score. The peak of activity can be found at 9 Kimura score having a genome proportion of 1.12% for DNA transposons, 0.02% for LTRs, 0.03% for LINEs, < 0.01 for SINEs, and 0.56% for unclassified TE sequences. MEAM1 has a peak of activity observed at 4 Kimura score having a genome proportion of 1.62% DNA transposons, 0.05% for LTRs, 0.06% for LINEs, 0.01 for

Table 4 Number of clusters shared across the three *B. tabaci* genomes

	DNA	LTR	LINE	SINE	Unknown	Retroelements	Total
ALL	216	174	120	2	220	1	733
MEAM1 and MED/Q	273	212	179	4	318	1	987
MEAM1 and SSA-ECA	133	74	56	0	283	0	546
MED/Q and SSA-ECA	126	74	56	0	183	2	441
Total	748	534	411	6	1004	4	2707

The table presents a summary of the number of clusters identified as shared across the *B. tabaci* genomes. The clusters were created from the TE consensus sequences from the six genomes included in the study. DNA represent DNA transposons, *LINE* Long interspersed nuclear elements, *SINE* Short interspersed nuclear elements, *LTR* Long terminal repeats

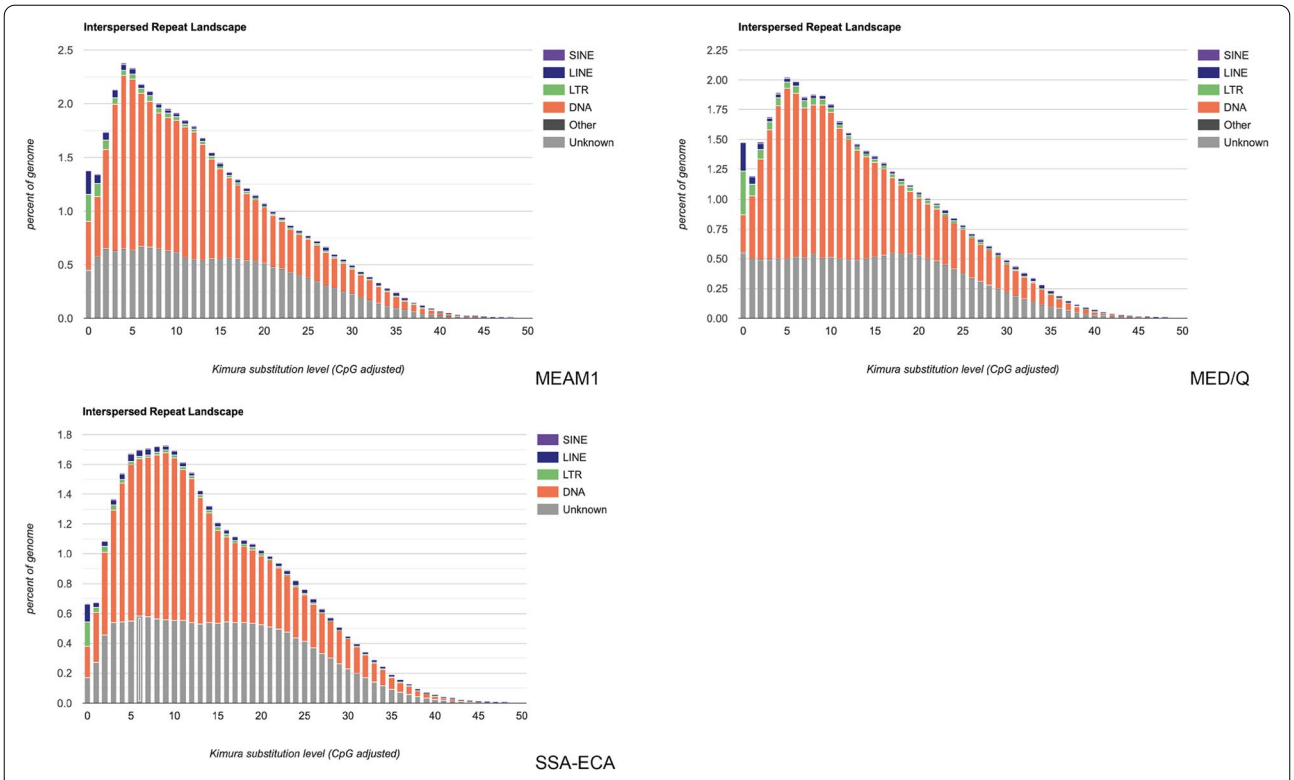

Fig. 4 Repeat landscapes of the *B. tabaci* genomes. The repeat landscapes illustrate the activity of the different classes of transposable elements found in the three *B. tabaci* genomes. Sequence divergence scores were measured using Kimura distance which is represented on the x-axis while the percent coverage of the element in the genome is represented on the y-axis. Elements with low sequence divergence scores represent a more recent transposable element activity while elements with higher sequence divergence scores represent older transposition events

SINEs, and 0.65% for unclassified TE sequences. MED/Q has a peak of activity at 5 Kimura score having a genome proportion of 1.43% for DNA transposons, 0.06% for LTRs, 0.03% for LINEs, 0.01% for SINEs, and 0.50% for unclassified TE sequences.

Further analysis of the superfamilies' activities reveals that majority of the DNA transposon superfamilies peaked at around the same timeline as the peak of their activity in each of the whitefly genomes while retrotransposons' peak of activity can be found at 0 Kimura score in all three landscapes (Supplementary Table 4).

Discussion

This study is the first to characterize TEs found within the *B. tabaci* species complex and create a standardized annotation workflow that could be used to analyse future whitefly genome releases. The first three publicly available genomes of the species complex were the focus of this analysis (MEAM1, MED/Q, and SSA-ECA). Our results highlight that previously published data suggesting there are marked differences in TE classes between species [14–16] is due to erroneous identifications of TEs in the MED/Q draft genome. The improved and standardized TE annotation workflow developed will allow a more accurate analysis of the distribution of TE across the whitefly species complex in future studies.

Identification of TEs in the genomes

The identification of TEs using the RepBase library yielded significantly lower results compared to the published results across the whitefly (Table 1) genomes while the RepBase library accurately identified TEs within the *D. melanogaster* genome (Table 2). In all three whitefly genomes, the TEs identified using the RepBase library were less than half of what was reported in their respective publications [14–16]. These results indicate that the RepBase library did not contain all the whitefly TE consensus sequences identified and published in respective previous studies [14–16]. Also as of April 12, 2019, RepBase is no longer publicly available and requires a subscription to access the up-to-date versions. These issues prevent further exploration of TEs within the species complex and have prompted the development of a TE annotation workflow that would standardize the annotation of multiple whitefly genomes.

The developed workflow was shown to accurately characterize TEs found within a genome using the *D. melanogaster* genome (Table 2). The repeats identified in the different *D. melanogaster* studies reported that < 20% of the genome is composed of TEs [51–54] and the results from the developed annotation workflow in this study were consistent with these findings. The similarities of the proportion of distribution of the TE orders shows the accuracy of the developed workflow. Research on *D. melanogaster* TEs date as far back as 1980 [59], and the TE annotation was able to identify these elements accurately.

This study attempted to run the TE annotation workflow described in the Chen et al. [14, 15] and Xie et al. [16] studies to compare results; however, the attempts did not yield similar results and some of the tools used failed to run with the other hemipteran genomes included in the study. The whitefly genome studies released their genome assemblies along with TE distribution content and GTF files for the TE copies; however, the TE consensus libraries were unavailable. These indicate that the TE libraries developed in the Chen et al. [14, 15] and Xie et al. [16] studies were not submitted to the RepBase library (or any other TE databases).

Within the whitefly genomes, the workflow developed was able to identify a similar proportion of TE orders within the MEAM1 and SSA-ECA genomes. Chen et al. [14, 15] reported the abundance of the DNA transposons in MEAM1 (29.25%) and SSA-ECA (25.94%) (Table 1). In contrast, in the MED/Q genome, LTRs were reported to be the most abundant element at 18.5%, with only 15.66% of the genome reported to be occupied by DNA transposons [16]. The results from this study show that the significant variation in the proportions of TEs found within these *B. tabaci* genomes is an artefact of the previous studies employing different TE annotation methods. In MEAM1 and SSA-ECA, a DNA transposon-specific identification tool was used while an LTR identification tool was included in the MED/Q annotation workflow. This study has highlighted the need for the implementation of a standardized workflow to accurately identify differences in TEs across genomes.

TE content and genome assembly size

A positive correlation between TE content and genome size has previously been reported in arthropod genomes [34, 60, 61] as well as other genomes [18, 44, 62]. An arthropod wide study conducted by Petersen et al. [34] was the most extensive showing the association of genome assembly size and TE proportion within arthropod genomes. The largest genome included in the Petersen et al. [34] study (*Locusta migratoria* 5759.8 Mbp) has the largest TE proportion (63.55% genome proportion) whilst the smallest genome studied (*Belgica antarctica* 89.54 Mbp) has the lowest TE proportion (2.58% genome proportion).

The same positive correlation was identified across the six genomes included in this study. The *B. tabaci* genomes on average were larger than the non-whitefly genomes and contain more TEs (Fig. 2). The *M. persicae* assembly, the smallest non-whitefly genome included in the study (347.31 Mbp), had the lowest TE content (17.52%). Although TE content within genomes has been consistently shown to correlate with genome size [34, 60, 63], it remains unclear as to how exactly TEs directly contribute to this as different arthropod genomes have different landscapes of TEs. In lepidopterans, TE length and activity have been linked to genome expansion; however, the exact order(s) of TEs which contributed to the expansion remains unclear [61]. An association of a specific TE order (DNA transposons) and genome assembly size was identified in the *Clitarchus hookeri* genome [60]; however, the extent of the relationship has not yet been fully explored.

TE classification

This study identified that the most abundant TE within the *B. tabaci* genomes are DNA transposons and are significantly higher within the whitefly species compared to the other hemipteran genomes included in the study. On average, the three whitefly genomes also had higher DNA transposons (22.85%) identified compared to the different arthropod clades that were analysed in the Petersen et al. study [34]; Hemiptera (3.24% average), Lepidoptera (1.40% average), Hymenoptera (2.83% average), and Drosophilids (1.67%).

DNA transposons are abundant in plant genomes and have been observed to have different roles; gene expression, genome expansion, gene regulation, and genome evolution [42, 60, 61]. DNA transposons can act as cis-regulatory elements which increase expression of nearby genes, and they can also decrease and silence gene expression because of small RNAs produced from them [41, 42].

In arthropods, DNA transposons have been observed to have a role in genome expansion [34, 60]. DNA transposons were identified to be the most abundant TE in the *C. hookeri* genome and comparison against the three other polyneopteran genomes shows an association of DNA transposons and genome assembly size [60]. The presence and absence of specific DNA TE superfamilies in the polyneopteran genomes have revealed the association; however, the mechanisms of the expansion due to the TEs require further exploration. The significant difference in DNA transposons found within the *B. tabaci* group could be one of the reasons why the genomes

found in the species complex are larger in size compared to the other hemipteran genomes included in the study.

The abundance of DNA transposons within the species complex has been reported in MEAM1 [14] and SSA-ECA [15] genomes but was not explored further. The presence of common and unique DNA transposon superfamilies across the whitefly genomes highlights the importance of this TE order within the species. A more exhaustive exploration beyond characterization would be required to further understand the context of the presence of these elements within the species complex.

There are significantly fewer LINEs in the whitefly compared to the three non-whitefly genomes studied. On average, the three whitefly genomes also had less LINEs (1.05%) identified compared to the different arthropod clades analysed in the Petersen et al. study [34]; Hemipterans (5.14% average), Lepidoptera (5.17% average), and Drosophilids (4.34%). Most LINE studies in insects have been done on drosophilids. In *D. melanogaster*, strains that carried specific non-LTR retrotransposons exhibited hybrid dysgenesis [28, 64]. The progenies of these insects became sterile and had an increase in the frequency of mutations and chromosome rearrangement [28, 64]. LINEs have been observed to successfully maintain themselves through their host organism's evolutionary lifetime [65–67]. Site-specific insertion of R1 and R2 LINE superfamilies near the 28S ribosomal RNA genes ensured its propagation while there is also evidence of another LINE superfamily successfully maintaining itself through non-site-specific insertion [65–68]. Different LINEs superfamilies can be found in the different insect genomes and each of these superfamilies could cause different effects depending on the type and the area of insertion [28, 34, 61, 69]. The consequences of the low distribution of LINEs within the whitefly species complex are unknown and an exploration of these elements in the wider context of insect evolution is warranted.

SINEs were the lowest identified TEs within the whitefly genomes. SINEs require LINEs for their transposition [28, 70] and the low distribution of SINEs within the species complex could correlate with the low distribution of LINEs. However, it should be noted that the workflow had difficulty identifying SINEs even when known SINEs were identified using the RepBase library (Table 1; Supplementary Table 1). The workflow was also unable to identify SINEs found within the *M. persicae* and *D. melanogaster* genomes.

The difficulty of identification of SINEs has been a consistent challenge in different arthropod TE studies. In the arthropod-wide TE identification performed by Petersen et al. [34], no SINE sequences were identified in seven of the 73 genomes included in the study. It is possible that there are genuinely no SINEs found within these genomes; however, there are multiple inconsistent reports of the proportions of identified SINEs found within the same organisms. Petersen et al. [34] reported 2.07% genome proport ion of SINEs within the *Heliconius melpomene* genome and 9.41% within the *B. mori* genome. In the *H. melpomene* TE analysis, Lavoie et al. [69] identified more at 8.22% genome proportion; whil e in the *B. mori* TE analysis done by Osanai-Futahashi et al. [71], 12.8% of the genome were identified as SINEs. The size of the retrotransposons adds to the difficulty of the identification of SINEs by automated TE annotation tools [72, 73]. SINEs being the shortest of the TEs would be impacted the most in the identification of these elements.

A significant percentage of TEs remain unclassified in the identified elements across genomes of the whitefly species complex. These unknown elements were screened for potential protein sequences and gene fragments; however, the screening yielded no positive results. Some of these unknown elements were also found to be shared across the three whitefly genomes. The significance of these unknown elements warrants further investigation and validation to enable improved classification and understanding of these elements.

Lastly, a third of the elements identified remain unknown, it should be noted that the distribution of classes amongst the TE class may change; however, DNA transposons in the *B. tabaci* species complex would remain the most abundant as more than half of the identified elements in the species complex are DNA transposons.

The landscape of TEs and the superfamilies within the *B. tabaci* genomes

The repeat landscapes highlight the difference of the TEs and their activity across the *B. tabaci* genomes. The most abundant DNA transposon superfamily across the three whitefly genomes is the hAT superfamily. The hAT superfamily represents one of the most well characterised transposable elements and it also includes the first mobile DNA element that was discovered which was the Activator maize transposon [45, 74]. The hAT superfamily's general structure is 2.5-5 k bp with terminal inverted repeats that could span up to 50 bp, generating up to 8 bp of target site duplication (TSD) and encoding a single protein that includes a transposase domain [27]. There are 13 additional hAT superfamilies representing distinct lineages that appeared in this study. Some specific elements in the hAT family have been explored to identify their functions, structure, and evolution [27, 74]. hAT-related sequences are found in different organisms, including humans, nematodes, flies, fungi, and plants [74].

For retrotransposons, there were three active repeat families found across the three *B. tabaci* genomes. Gypsy and Pao were found to be the most active LTR superfamily in all th ree genomes. Gypsy elements were first characterised in the *D. melanogaster* genome and their sequences have a high similarity with retroviruses of vertebrate animals [28]. Gypsy elements have high rates of transposition and are shown to insert themselves in introns and affect gene expression by disrupting normal transcriptional control [75, 76]. Pao elements are LTR elements that are related to the Gypsy element and are said to originate from the *Bombyx mori* genome [28, 77]. Pao elements encode a GAG and pol proteins and create a 4–6 TSD once they are inserted in the genome [25, 77]. In LINEs, RTE-BovB is the most active superfamily in the *B. tabaci* genomes. In the RepBase classification, BovB is classified under the RTE group where repeats in this group have the ability to encode their protein with two functioning domains; AP-endonuclease (Apurinic) and a reverse transcriptase [25, 27]. Bov-B (Bovine-B) elements have been identified in the *Bos taurus* genome and have been observed to have horizontal transfer events in other eukaryotic genomes [78–81].

The shape of the distribution of the repeats is similar within MEAM1 and MED/Q genomes. MEAM1 and MED/Q also share the greatest number of clusters. These two whitefly species are relatively closely related in phylogenetic analyses of the *B. tabaci* species complex [8, 82]. Aside from the shape of the distribution, the trends in the expansion are also similar between the two genomes as they both have the same superfamilies that are currently the most active, namely CACTA, hAT, RTE-BovB, Copia, Pao, and Gypsy. Copia's activity was more prominent in the *B. tabaci* group with its presence being at low genome coverage in the non-whitefly group with the exception of the *A. pisum* genome assembly. Copia elements are autonomous LTR retrotransposons and their defining feature is the position of the integrase domain [27, 83]. Copia elements can be traced back further in plants while found to be more recently active in insects [84, 85]. They have been recently active in the *Drosophila* genome and it is hypothesized that they may be horizontally transmitted [85].

There is a decrease in the expansion of the activity of DNA transposon and an increase in LINE and LTR activity in the *B. tabaci* genome assemblies. It is still not fully clear how these trends affect their respective genomes. The relative age of the elements was identified using the Kimura substitution model; however, in order to place the element's age within a more objective timescale, there is a need to determine the rate of evolution in whiteflies.

Future of TE research in the whitefly species complex

With the availability of a standardized workflow and characterized TEs within the whitefly species complex, further investigation of the activity of these elements can now be performed. The impact that TEs have on biological properties (e.g., host plant colonisation, polyphagy, detoxification, virus transmission) and diversification of members of the whitefly species complex would be priority areas for further studies.

Conclusion

TEs occupy a significant portion of whitefly genomes yet to date there have been no studies that characterise accurately the distribution of TEs found within the *B. tabaci* species complex. This study is the first to explore TE distribution within the species complex and to create a workflow to standardize the characterization of the elements across multiple whitefly genomes. The standardization of the TE annotation workflow has identified an abundance of DNA transposons within the species complex and has shown this to be true across all published *B. tabaci* genomes, contradicting previously published results [16]. Other TE superfamilies of note were also identified, some of these superfamilies were shown to be specific to the whitefly genomes. Unclassified elements remain significant, and the biological implications of the known elements also remain unknown. These issues highlight the need to explore further these elements within the different genomes of this whitefly species complex. The study has provided the groundwork for future TE studies within the species and hopes the initial characterization of these elements will increase interest in TEs found within the *B. tabaci* species complex.

Methodology

Genome data sets

Six different arthropod genomes were included in the study. Three of the genomes are from the *B. tabaci* cryptic species complex were included in the study; MEAM1 [14], MED/Q [16], and SSA-ECA [15]. The MEAM1 (*B. tabaci* Middle East-Asia Minor 1) genome assembly was obtained from GenBank under the accession number GCA_001854935.1. The MED/Q (*B. tabaci* Mediterranean) genome assembly was obtained from www.gigadb.org/dataset/100286. The SSA1-ECA (Sub-Saharan 1 East Central Africa) genome assembly was obtained from ftp://www.whiteflygenomics.org/pub/whitefly/SSA-ECA/v1.0/.

The three other arthropod genomes were non-whitefly genomes and were included to assess the performance of

the workflow and compare the results of the TEs identified with the whitefly genomes; *Acyrthosiphon pisum* (project accession ABLF01000000)[86], *Diaphorina citri* (project accession AWGM01000000)[87], and *Myzus persicae* (project code LXJY01000000)[88]. All three genome assemblies were obtained from NCBI using their project accession codes.

Repeat identification

The workflow performed in this study for creating a species-specific repeat library. The genome assembly to be studied is first submitted to MITETracker (https://github.com/INTABiotechMJ/MITE-Tracker) [89] and TransposonPSI_08222010 (http://transposonpsi.sourceforge.net/) for the initial step of the identification. The genome assembly was then submitted to genometools v1.5.9 (LTRHarvest and LTRDigest). Elements ranging from 100 to 6000 bps with terminal ending repeats with ≥ 85% similarity are identified as LTRs. The TEs representative sequences produced from MITETracker and genometools are then combined to create one library and is submitted to RepeatMasker to mask copies of the TEs found in the genome assembly. The masked genome assembly is then submitted to RepeatModeler v1.0.11 [90] for de novo TE identification. The masking of the copies of the already identified TE copies prevents RepeatModeler from identifying and modelling the repeat sequences that have already been identified. Utility scripts from the MAKER-P pipeline were also used to aid with the parsing of the results of genometools v1.5.9 (LTRHarvest and LTRDigest), RepeatModeler v1.0.11, and RepeatMasker v4.1.1 [91].

Each of the programs has candidate sequences that they have identified as repeat elements and the four outputs are subsequently merged into one library that is then submitted to USEARCH v11.0.667 [92, 93]. The clustering algorithm by USEARCH utilizes a algorithm called "greedy algorithm" which implements the "best" solution based on the current options. This means that sequence input order is important in the identification of candidate consensus sequences as the options for each cluster is based on the order of the sequences in the library. Sorting was performed using USEARCH's "-sortbylength" command and the clusters were created based on ≥ 80% sequence similarity. A consensus sequence is then produced from each of the clusters to obtain a representative sequence. The representative sequences have ≥ 80% similarity to the member sequences. All the representative sequences have < 80% similarity to each other. The process reduces redundancy and assists in the identification of degenerated repeat elements.

The repeat library produced by the repeat identification workflow underwent several series of steps to classify each of the consensus sequences. The first method used for classification was the homology-based approach. The repeat library is submitted to RepeatClassifier(https://github.com/rmhubley/RepeatModeler/blob/master/RepeatClassifier) and the unclassified sequences were subsequently submitted to the web browser version of Censor [94]. Before continuing to the next step of the classification, sequences < 70 bp were removed and the sequences which were classified by the methods. The library was then submitted to TEClass v2.1.3 [95] and PASTEClassifier v1.0 [96]. Manual curation was done to analyse the results of both tools. The curation was based on sequence similarity and the length of the sequence aligned. The classification was accepted when both tools had similar results and spanned ≥ 80% of the element's length. When the results from the classification differ in the class level (i.e. DNA transposon and Retrotransposon), the element remained unknown. When the classification resulted in a difference in order (i.e. LINEs vs SINEs, LTR vs NonLTR) and ≥ 80% of the sequence length was identified, the element was classified using its more general level of classification. Any results that had less than 80% sequence length was disregarded..

A Blast search was performed on the unknown sequences against the NCBI nr protein database (version 2019.08.05) and UniProtKB/Swiss-Prot Arthropoda protein sequences obtained on July 10, 2019. The plan was to identify the unknown sequences with hits and parse through the results and remove the sequences with more than 50 bps hits from the species-specific library. None of the unknown sequences yielded a blast result and the unknown sequences were accepted as unclassified TEs.

Results from the homology-based classification, the consensus classification of TEClass and PASTEClassifier, and the unknown sequences were then combined to produce the final library. The process was repeated for each of the repeat libraries produced from the genomes included in this study.

Genome assembly size and TE Distribution across species analysis

The proportion of TEs found within each genome were obtained from the RepeatMasker v4.1.1 output tables. The relationship between genome assembly size and TE content across the six genomes was tested using Spearman's rank rho correlation. Spearman rank correlation tests the association between either two rank variables or one ranked and one measurement variable. The relationship identifies whether the variables covary (the variable increases/decreases when the other variable's value changes).

A standard t-test and Wilcoxon rank-sum test were used to further compare the proportion of each order TEs across each group of genomes. Both tests compare the mean values of a measurement variable and identify if the mean values are significantly different. In this study, the tests identified whether there is a significant difference between the TE proportion of each order between the whitefly and non-whitefly genomes. The standard t-test was used for the TE orders with a similar variance while the Wilcoxon rank-sum test was used for values with non-normal distribution.

Abbreviations

BAC: Bacterial Artificial Chromosome; GTF: Gene transfer format; LINE: Long interspersed nuclear elements; LTR: Long terminal repeat; MITE: Miniature inverted repeat transposable element; NonLTR: Non-long terminal repeat; SINE: Short interspersed nuclear elements; TE: Transposable element.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13100-022-00270-6.

Additional file 1. Supplementary Table 1.

Additional file 2. Supplementary Table 2.

Additional file 3. Supplementary Table 3.

Additional file 4. Supplementary Table 4.

Acknowledgements

The authors would like to thank the University of Greenwich for a Vice Chancellor Scholarship to JPS. This work was also supported in part by the Bill & Melinda Gates Foundation [Grant Number OPP1149777]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Authors' contributions

JPAS gathered, analysed, and interpreted the data used in this study. JPAS and SOS developed and tested the TE identification workflow. PV verified the methods used in the study. SES, PV and SB provided supervision to JPAS. JPAS drafted the manuscript and all authors contributed to the editing of the final manuscript. The author(s) read and approved the final manuscript.

Funding

JPS was funded through a Vice Chancellor Scholarship from the University of Greenwich. Contributions from PV, SOS, SB and SS were supported in part through the University of Greenwich and in part through the Bill & Melinda Gates Foundation [Grant Number OPP1149777].

Availability of data and materials

MEAM1, *A. pisum*, *D. citri*, and *M. persicae* genome assemblies are available at NCBI under the accession number GCA_001854935.1, project code ABLF01000000, project code AWGM01000000, and project code LXJY01000000.MED/Q genome assembly is available at www.gigadb.org/dataset/100286. SSA1-ECA genome assembly is available at ftp://www.whiteflygenomics.org/pub/whitefly/SSA-ECA/v1.0/. The species–specific repeat libraries have been submitted to DFAM and is currently under review. The species-specific repeat libraries and the associated downstream analysis script are also available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors have no competing interest to declare.

Author details

[1]Natural Resources Institute, University of Greenwich, Central Avenue, Gillingham, Chatham ME4 4TB, UK. [2]Centre for Agriculture and the Bioeconomy, Queensland University of Technology, Brisbane, QLD 4000, Australia.

Received: 30 January 2022 Accepted: 30 March 2022
Published: 19 April 2022

References

1. Seal SE, VandenBosch F, Jeger MJ. Factors influencing begomovirus evolution and their increasing global significance: Implications for sustainable control. CRC Crit Rev Plant Sci. 2006;25:23–46. https://doi.org/10.1080/07352680500365257.
2. Naranjo SE, Chu CC, Henneberry TJ. Economic injury levels for Bemisia tabaci (Homoptera: Aleyrodidae) in cotton: Impact of crop price, control costs, and efficacy of control. Crop Prot. 1996;15:779–88. https://doi.org/10.1016/s0261-2194(96)00061-0.
3. Oliveira MRV, Henneberry TJ, Anderson P. History, current status, and collaborative research projects for Bemisia tabaci. Crop Prot. 2001;20:709–23.
4. Martin JH, Mound LA. An annotated check list of the world's whiteflies. Magnolia Press; 2007. www.mapress.com/zootaxa/.
5. Abd-Rabou S, Simmons AM. Survey of reproductive host plants of Bemisia tabaci (Hemiptera: Aleyrodidae) in Egypt, including new host records. Entomol News. 2010;121:456–65. https://doi.org/10.3157/021.121.0507.
6. Navas-Castillo J, Fiallo-Olivé E, Sánchez-Campos S. Emerging virus diseases transmitted by whiteflies. Annu Rev Phytopathol. 2011;49:219–48.
7. MacFadyen S, Paull C, Boykin LM, De Barro P, Maruthi MN, Otim M, et al. Cassava whitefly, Bemisia tabaci (Gennadius) (Hemiptera: Aleyrodidae) in East African farming landscapes: A review of the factors determining abundance. Bull Entomol Res. 2018;108:565–82. https://doi.org/10.1017/S0007485318000032.
8. Mugerwa H, Colvin J, Alicai T, Omongo CA, Kabaalu R, Visendi P, et al. Genetic diversity of whitefly (Bemisia spp.) on crop and uncultivated plants in Uganda: implications for the control of this devastating pest species complex in Africa. J Pest Sci. 2021;94:1307–30. https://doi.org/10.1007/s10340-021-01355-6.
9. Malka O, Santos-Garcia D, Feldmesser E, Sharon E, Krause-Sakate R, Delatte H, et al. Species-complex diversification and host-plant associations in Bemisia tabaci: A plant-defence, detoxification perspective revealed by RNA-Seq analyses. Mol Ecol. 2018;27:4241–56. https://doi.org/10.1111/mec.14865.
10. Malka O, Feldmesser E, van Brunschot S, Santos-Garcia D, Han WH, Seal S, et al. The molecular mechanisms that determine different degrees of polyphagy in the Bemisia tabaci species complex. Evol Appl. 2021;14:807–20. https://doi.org/10.1111/eva.13162.
11. Aidlin Harari O, Santos-Garcia D, Musseri M, Moshitzky P, Patel M, Visendi P, et al. Molecular Evolution of the Glutathione S-Transferase Family in the Bemisia tabaci Species Complex. Genome Biol Evol. 2020;12:3857–72.
12. Chi Y, Pan LL, Bouvaine S, Fan YY, Liu YQ, Liu SS, et al. Differential transmission of Sri Lankan cassava mosaic virus by three cryptic species of the whitefly Bemisia tabaci complex. Virology. 2020;540:141–9. https://doi.org/10.1016/j.virol.2019.11.013.
13. Fan YY, Zhong YW, Zhao J, Chi Y, Bouvaine S, Liu SS, et al. Bemisia tabaci vesicle-associated membrane protein 2 interacts with begomoviruses and plays a role in virus acquisition. Cells. 2021;10(7):1700.
14. Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan J, et al. The draft genome of whitefly Bemisia tabaci MEAM1, a global crop pest,

provides novel insights into virus transmission, host adaptation, and insecticide resistance. BMC Biol. 2016;14:110. https://doi.org/10.1186/s12915-016-0321-y.
15. Chen W, Wosula EN, Hasegawa DK, Casinga C, Shirima RR, Fiaboe KKM, et al. Genome of the African cassava whitefly Bemisia tabaci and distribution and genetic diversity of cassava-colonizing whiteflies in Africa. Insect Biochem Mol Biol. 2019;110:112–20. https://doi.org/10.1016/j.ibmb.2019.05.003.
16. Xie W, Chen C, Yang Z, Guo L, Yang X, Wang D, et al. Genome sequencing of the sweetpotato whitefly Bemisia tabaci MED/Q. Gigascience. 2017;6:1–7. https://doi.org/10.1093/gigascience/gix018.
17 Correa M, Lerat E, Birmelé E, Samson F, Bouillon B, Normand K, et al. The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality. Genome Biol Evol. 2021;13:eva062. https://doi.org/10.1093/gbe/evab062.
18. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115:49–63. https://doi.org/10.1023/A:1016072014259.
19. Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, Myers EW, et al. Improved repeat identification and masking in Dipterans. Gene. 2007;389:1–9. https://doi.org/10.1016/j.gene.2006.09.011.
20. Holt C, Yandell M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491. https://doi.org/10.1186/1471-2105-12-491.
21. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015;16:184. https://doi.org/10.1186/s13059-015-0729-7.
22 Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989;5 C:103–7.
23. Finnegan DJ. Transposable elements. Curr Opin Genet Dev. 1992;2:861–7.
24. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
25. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
26. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol. 2015;86:90–109.
27. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst. 2019;94:233–52.
28. Galun E. Transposable Elements. Dordrecht: Springer Netherlands; 2003. https://doi.org/10.1007/978-94-017-3582-7.
29. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331–68. https://doi.org/10.1146/annurev.genet.40.110405.090448.
30. Eickbush TH. Retrotransposons. In: Brenner S, Miller JHBT-E of G, editors. Encyclopedia of Genetics. New York: Academic Press; 2001. p. 1699–701. https://doi.org/10.1006/rwgn.2001.1111.
31. Eickbush TH, Malik HS. Origins and Evolution of Retrotransposons. In: Mobile DNA II. American Society of Microbiology; 2014. p. 1111–44.
32. Finnegan DJ. Retrotransposons. Curr Biol. 2012;22:R432–7. https://doi.org/10.1016/j.cub.2012.04.025.
33. Kazazian HH, Scott AF. "Copy and paste" transposable elements in the human genome. J Clin Invest. 1993;91:1859–60. https://doi.org/10.1172/JCI116400.
34. Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol Biol. 2019;19:11. https://doi.org/10.1186/s12862-018-1324-9.
35. Parisot N, Vargas-Chávez C, Goubert C, Baa-Puyoulet P, Balmand S, Beranger L, et al. The transposable element-rich genome of the cereal pest Sitophilus oryzae. BMC Biol. 2021;19:2021.03.03.408021. https://doi.org/10.1186/s12915-021-01158-2.
36. Griffiths a. JF, Gelbart WM, Lewontin RC, Miller JH. Modern Genetic Analysis. Second. New York: W.H. Freeman & Co. Ltd; 2002. 2020–03–23.
37. Gilbert C, Peccoud J, Cordaux R. Transposable Elements and the Evolution of Insects. Annu Rev Entomol. 2021;66:355–72. https://doi.org/10.1146/annurev-ento-070720-074650.
38. Hiraizumi Y. Spontaneous recombination in Drosophila melanogaster males. Proc Natl Acad Sci U S A. 1971;68:268–70. https://doi.org/10.1073/pnas.68.2.268.
39. Majumdar* S, Rio DC. P Transposable Elements in Drosophila and other Eukaryotic Organisms . Microbiol Spectr. 2015;3:MDNA3–2014. https://doi.org/10.1128/microbiolspec.mdna3-0004-2014.
40. Kelleher ES. Reexamining the P-element invasion of Drosophila melanogaster through the lens of piRNA silencing. Genetics. 2016;203:1513–31. https://doi.org/10.1534/genetics.115.184119.
41. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009;461:1130–4.
42 Han MJ, Zhou QZ, Zhang HH, Tong X, Lu C, Zhang Z, et al. IMITEdb: The genome-wide landscape of miniature inverted-repeat transposable elements in insects. Database. 2016;2016:baw48. https://doi.org/10.1093/database/baw148.
43. Kim J, Martignetti JA, Shen MR, Brosius J, Deininger P. Rodent BC1 RNA gene as a master gene for ID element amplification. Proc Natl Acad Sci U S A. 1994;91:3607–11. https://doi.org/10.1073/pnas.91.9.3607.
44. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements 06 Biological Sciences 0604 Genetics. Genome Biol. 2018;19:199. https://doi.org/10.1186/s13059-018-1577-z.
45. McCLINTOCK B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36:344–55. https://doi.org/10.1073/pnas.36.6.344.
46. Biémont C. A brief history of the status of transposable elements: From junk DNA to major players in evolution. Genetics. 2010;186:1085–93. https://doi.org/10.1534/genetics.110.124180.
47. Cerbin S, Jiang N. Duplication of host genes by transposable elements. Curr Opin Genet Dev. 2018;49:63–9. https://doi.org/10.1016/j.gde.2018.03.005.
48. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 2005;37:997–1002.
49. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65:505–30.
50. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 2015;25:445–58. https://doi.org/10.1101/gr.185579.114.
51. Goubert C, Modolo L, Vieira C, Moro CV, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol. 2015;7:1192–205. https://doi.org/10.1093/gbe/evv050.
52. Tom Hill. Transposable element dynamics are consistent across the Drosophila phylogeny, despite drastically differing content. bioRxiv. 2019;2:1–29. https://doi.org/10.1101/651059.
53. Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable elements in Drosophila. Mob. DNA. 2020;11:23. https://doi.org/10.1186/s13100-020-00213-z.
54. Repeatmasker.org. D. melanogaster [Drosophila melanogaster] Genomic Data set. http://www.repeatmasker.org/species/dm.html. Accessed 12 Jan 2020.
55 Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 2002;3:RESEARCH0084-RESEARCH0084. https://doi.org/10.1186/gb-2002-3-12-research0084.
56. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
57. Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A. 2011;108:7884–9. https://doi.org/10.1073/pnas.1104208108.
58. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.
59 Bregliano JC, Picard G, Bucheton A, Pelisson A, Lavige JM, L'Heritier P. Hybrid dysgenesis in drosophila melanogaster. Science (80-). 1980;207:606–11. https://doi.org/10.1126/science.6766221.

60. Wu C, Twort VG, Crowhurst RN, Newcomb RD, Buckley TR. Assembling large genomes: Analysis of the stick insect (Clitarchus hookeri) genome reveals a high repeat content and sex-biased genes associated with reproduction. BMC Genomics. 2017;18:884. https://doi.org/10.1186/s12864-017-4245-x.
61. Talla V, Suh A, Kalsoom F, Dinca V, Vila R, Friberg M, et al. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (leptidea) butterflies. Genome Biol Evol. 2017;9:2491–505. https://doi.org/10.1093/gbe/evx163.
62. Naville M, Henriet S, Warren I, Sumic S, Reeve M, Volff JN, et al. Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. Curr Biol. 2019;29:1161-1168.e6.
63. Sessegolo C, Burlet N, Haudry A. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol Lett. 2016;12:20160407. https://doi.org/10.1098/rsbl.2016.0407.
64 Fawcett DH, Lister CK, Kellett E, Finnegan DJ. Transposable elements controlling I-R hybrid dysgenesis in D. melanogaster are similar to mammalian LINEs. Cell. 1986;47:1007–15.
65. Lathe WC, Burke WD, Eickbush DG, Eickbush TH. Evolutionary stability of the R1 retrotransposable element in the genus Drosophila. Mol Biol Evol. 1995;12:1094–105.
66. Lathe WC, Eickbush TH. A single lineage of R2 retrotransposable elements is an active, evolutionarily stable component of the Drosophila rDNA locus. Mol Biol Evol. 1997;14:1232–41. https://doi.org/10.1093/oxfordjournals.molbev.a025732.
67. Biedler JK, Tu Z. The Juan non-LTR retrotransposon in mosquitoes: Genomic impact, vertical transmission and indications of recent and widespread activity. BMC Evol Biol. 2007;7:112. https://doi.org/10.1186/1471-2148-7-112.
68. Jakubczak JL, Burke WD, Eickbush TH. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. Proc Natl Acad Sci U S A. 1991;88:3295–9. https://doi.org/10.1073/pnas.88.8.3295.
69. Lavoie CA, Platt RN, Novick PA, Counterman BA, Ray DA. Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera. Mob DNA. 2013;4:21. https://doi.org/10.1186/1759-8753-4-21.
70. Ohshima K, Okada N. SINEs and LINEs: Symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res. 2005;110:475–90. https://doi.org/10.1159/000084981.
71. Osanai-Futahashi M, Suetsugu Y, Mita K, Fujiwara H. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm. Bombyx mori Insect Biochem Mol Biol. 2008;38:1046–57. https://doi.org/10.1016/j.ibmb.2008.05.012.
72. Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, et al. Classification and characterization of human endogenous retroviruses mosaic forms are common. Retrovirology. 2016;13:7. https://doi.org/10.1186/s12977-015-0232-y.
73. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7. https://doi.org/10.1073/pnas.1921046117.
74. Rubin E, Lithwick G, Levy AA. Structure and evolution of the hAT transposon superfamily. Genetics. 2001;158:949–57. https://doi.org/10.1093/genetics/158.3.949.
75. Herédia F, Loreto ELS, Valente VLS. Complex evolution of gypsy in drosophilid species. Mol Biol Evol. 2004;21:1831–42. https://doi.org/10.1093/molbev/msh183.
76 Kim A, Terzian C, Santamaria P, Pélisson A, Prud'homme N, Bucheton A. Retroviruses in invertebrates: The gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. Proc Natl Acad Sci U S A. 1994;91:1285–9. https://doi.org/10.1073/pnas.91.4.1285.
77. Xiong Y, Burke WD, Eickbush TH. Pao, a highly divergent retrotransposable element from Bombyx mori containing long terminal repeats with tandem copies of the putative R region. Nucleic Acids Res. 1993;21:2117–23. https://doi.org/10.1093/nar/21.9.2117.
78. Peccoud J, Loiseau V, Cordaux R, Gilbert C. Massive horizontal transfer of transposable elements in insects. Proc Natl Acad Sci U S A. 2017;114:4721–6. https://doi.org/10.1073/pnas.1621178114.
79. Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL. Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci U S A. 2013;110:1012–6. https://doi.org/10.1073/pnas.1205856110.
80 Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, et al. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science (80-). 2009;324:522–8. https://doi.org/10.1126/science.1169588.
81. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 2018;19:85. https://doi.org/10.1186/s13059-018-1456-7.
82. De Barro PJ, Liu SS, Boykin LM, Dinsdale AB. Bemisia tabaci: A statement of species status. Annu Rev Entomol. 2011;56:1–19.
83. Qiu F, Ungerer MC. Genomic abundance and transcriptional activity of diverse gypsy and copia long terminal repeat retrotransposons in three wild sunflower species. BMC Plant Biol. 2018;18:6. https://doi.org/10.1186/s12870-017-1223-z.
84. Sabot F, Schulman AH. Parasitism and the retrotransposon life cycle in plants: A hitchhiker's guide to the genome. Heredity (Edinb). 2006;97:381–8. https://doi.org/10.1038/sj.hdy.6800903.
85. White SE, Habera LF, Wessler SR. Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci U S A. 1994;91:11792–6. https://doi.org/10.1073/pnas.91.25.11792.
86. Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D, et al. Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol. 2010;8(2):e10000313.
87 Saha S, Hosmani PS, Villalobos-Ayala K, Miller S, Shippy T, Flores M, et al. Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. Database. 2017;2017:bax032. https://doi.org/10.1093/database/bax032.
88. Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, et al. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. Genome Biol. 2017;18:27. https://doi.org/10.1186/s13059-016-1145-3.
89. Crescente JM, Zavallo D, Helguera M, Vanzetti LS. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics. 2018;19:348. https://doi.org/10.1186/s12859-018-2376-y.
90. Smit A, Hubley R. RepeatModeler Open-1.0. 2008. http://www.repeatmasker.org.
91. Campbell MS, Law MY, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: A Tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164:513–24. https://doi.org/10.1104/pp.113.230144.
92. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1. https://doi.org/10.1093/bioinformatics/btq461.
93. Edgar RC. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10:996–8. https://doi.org/10.1038/nmeth.2604.
94. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 2006;7:474. https://doi.org/10.1186/1471-2105-7-474.
95. Abrusán G, Grundmann N, Demester L, Makalowski W. TEclass - A tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics. 2009;25:1329–30.
96. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An automatic transposable element classification tool. PLoS ONE. 2014;9: e91929. https://doi.org/10.1371/journal.pone.0091929.

Publisher's Note