Characterization of transposable elements within the Bemisia tabaci species complex

Juan Paolo A. Sicat (✉ J.A.Sicat@greenwich.ac.uk)

University of Greenwich

Paul Visendi Muhindira

Queensland University of Technology

Steven O Sewe

University of Greenwich

Sophie Bouvaine

University of Greenwich

Susan E Seal

University of Greenwich

Research Article

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

Posted Date: February 2nd, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1312818/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at Mobile DNA on April 19th, 2022. See the published version at <https://doi.org/10.1186/s13100-022-00270-6>.

Abstract

Background

Whiteflies are agricultural pests that have caused worldwide negative impacts that have led to severe financial losses. The *Bemisia tabaci* whitefly species complex is the most damaging in terms of their broad crop host range and its ability to serve as vector for over 300 plant viruses. Whitefly genomes of the species complex provide valuable genomic data; however, transposable elements (TEs) within the species complex remain unexplored. This study provides the first accurate exploration of TE content within the *B. tabaci* species complex.

Results

This study identified an average of 40.61% of the genomes of three whitefly species (MEAM1, MEDQ, and SSA-ECA) consists of TEs. 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 the three whitefly genomes with three other hemipteran genomes and found a significant difference in the presence of DNA transposons and LINEs. 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) of which 11 TE superfamilies were identified to not be present in the three other hemipteran genomes (nine DNA transposon, and two LINE).

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* species complex. The characterization of these elements within the three whitefly genomes shows that TEs occupy a significant portion of the whitefly genome, majority of which are DNA transposons. This study also identified TE superfamilies of note and provides a framework for future TE studies within the species complex.

Introduction

Whiteflies are agricultural pests that cause crop losses that have a worldwide agricultural impact and have led to millions of dollars in financial losses [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 >300 plant viruses [4–6].

Agricultural intensification, climate change, and trade are exacerbating the global dispersal leading to super-abundant populations of *B. tabaci* [1, 7, 8]. The severity of this pest species complex has shaped several national and international collaborative projects [3], leading to an increase in genome and transcriptome resources. These will assist 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 sequences 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 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 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 antarctica* to as high as 72.8% in *Sitophilus oryzae* [34, 35]. The functions of these elements are often unknown, but their presence in genomes have 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 role of TEs in gene regulation and expression have already been established [28, 41–46] and the abundance of the element in the different whitefly genomes could play an important role in the evolution of the species complex. TEs have also been associated with gene duplication wherein the insertion location of the TE affect 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 have been characterized and explored in different arthropod genomes; however, TEs remain unexplored within the *B. tabaci* species complex.

TEs represent a major proportion of *B. tabaci* genomes, accounting for approximately 40–44% of the complete *B. tabaci* draft genomes published (MEAM1 and MED/Q) [14, 16]. The latest released *B. tabaci* genome, SSA-ECA, was incomplete and reported a slightly lower (38.5%) TE content but noted that around a quarter of genome data was missing from their 513 Mb genome assembly [15] and hence the repeat content 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, published studies have reported estimates of TE orders in specific *B. tabaci* genomes that when compared across studies show marked differences in TE orders within the *B. tabaci* complex. Although all the studies reported having around ~40% of the genome comprise of TEs, the MEAM1 and SSA-ECA whitefly genomes were reported to have an abundance of DNA transposons particularly MITEs (miniature inverted-repeat transposable elements) while LTRs were reported 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 are considered potentially significant.

The studies reporting very different TE class proportions in the *B. tabaci* whitefly genomes employed different TE annotation workflows. In both MEAM1 and SSA-ECA annotation, 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. 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 use of different workflows hinders the accurate comparisons of TE classes across the three *B. tabaci* genomes. Reliable inferences based on significant differences in TE compositions found across 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 published estimates using the same genome assemblies. To address this issue, 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 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 published genomes from the *B. tabaci* cryptic species complex were the focus of the analyses (MEAM1, MED/Q, and SSA-ECA). 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 shows 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. There was an attempt to find the rest of the consensus sequences; however, publication of these consensus sequences could not be found.

The RepBase library was then 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 applied more widely. The RepBase library was able to identify 17.44% TE genome coverage 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

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.

	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

The results of the TE identification using the RepeatMasker RepBase library shows 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 needed to be developed to standardize the identification of the TEs across the whitefly genomes. The different whitefly genomes published utilized different TE identification tools; MEAM1, and SSA-ECA used a DNA transposons specific tool, while MED/Q used a LTR specific identification tool. Standardization of the annotation workflow would allow a fairer comparison across the three genomes. A species-specific custom-built repeat library was created for each genome studied using the same range of

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 coverage of interspersed repeats was found in the *D. melanogaster* using the RepeatMasker library compared to 16.88% genome coverage 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 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

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.

	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

TEs in arthropod genomes

The developed workflow after validation was used to identify the TE content of each of the target genomes (Figure 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 hemipteran genomes were included, namely *Acyrtosiphon pisum* (ACPIS), *Diaphorina citri* (DIPSY), and *Myzus persicae* (MYPER). Each of the three whitefly genomes had a higher TE content (an average of 40.61% genome coverage of TEs) compared to each of the three non-whitefly genomes (an average of 25.01% TE genome coverage). MEAM1 had the highest TE content across the six genomes at 44.14% while ACPIS had the highest TE content amongst the non-whitefly genomes at 34.54%. SSA-ECA had the lowest TE content amongst the whitefly genomes at 36.80% but was still higher than the TE content in the ACPIS genome. MYPER had the lowest TE content across the six genomes at 17.52%.

The relationship between genome sizes of the six genomes and their TE content was tested using Spearman's rank rho correlation (Figure 2). TE coverage was found to be positively correlated with genome 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, MYPER (347 Mbp) had the lowest TE content at 17.52%. Amongst the whitefly genomes, SSA-ECA has the smallest genome size (538.48 Mbp) and the lowest TE genome coverage (36.80%).

Difference in the distribution of TE content between genomes

There was no statistically significant difference ($p = 0.09$) in genome size between the whitefly genomes (average 603.92 Mbp) and the non-whitefly genomes (average 458.24 Mbp). This allows us to compare the two groups without significantly biasing our results with the variations in genome 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% coverage 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 ACPIS (14.06%) and MYPER (8.35%) while retrotransposons were the most abundant class in the DIPSY genome (6.68%). An average of 4.34% coverage was identified as retrotransposons within the non-whitefly genomes. LINEs were the most abundant retrotransposon order in ACPIS (2.32%) and MYPER (1.86%) while SINEs were the most abundant in DIPSY (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, DIPSY had the highest percentage of SINEs at 3% while this TE order was not detected in MYPER.

The distribution of TEs between the genomes was explored further by comparing their distribution between the two groups of genomes. 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) (Figure 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 coverage 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.

Lastly, unclassified elements are still found within the identified TEs. Across the six genomes, an average of 13.70% genome coverage remains unclassified (15.43% for the whitefly genomes and 11.98% for the non-whitefly genomes). The relative proportions of the elements will therefore be subject to change when these 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 MYPER has 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 DIPSY had the least at 30 superfamilies. SINE superfamilies were the least identified at an average of four superfamilies. DIPSY had the greatest number of SINE superfamilies identified with seven while SINEs were not identified at all in MYPER.

Table 3

Repeat Superfamilies identified within the genomes

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).

	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
ACPIS	43	18	5	1	67
DIPSY	30	23	6	7	66
MYPER	36	18	7	0	61

TE repeat superfamilies that had been identified across the three whitefly genomes were analyzed further (Figure 4A). A total of 63 superfamilies were found to be common across the three whitefly genomes, 39 of these superfamilies were identified as DNA transposons, six LTR, 16 LINEs, and two SINEs. Aside from the common superfamilies, each genome had repeat superfamilies that were identified uniquely in them. MED/Q had the highest number of unique superfamilies at ten consisting of six DNA transposons (hAT-hAT19, Kolobok-E, Kolobok-T2, PIF-ISL2EU, TcMar, and TcMar-Sagan), two LINEs (CR1-Zenon and Daphne) and two SINEs (SINE2 and tRNA-V). Seven unique superfamilies were identified in the MEAM1 genome consisting of four DNA transposon superfamilies (Crypton-S, hAT-hAT1, P-Fungi, and TcMar-Cweed), two LTR superfamilies (ERVL and Caulimovirus), and one SINE superfamily (tRNA-L2). SSA-ECA had the least number of unique superfamilies identified at six which consisted of four DNA transposon superfamilies (hAT-hATw, IS, TcMar-ISRm11, and TcMar-Stowaway) and two LTR superfamilies (DIRS and Foamy).

Repeat superfamilies identified across the three non-whitefly genomes were also analyzed (Figure 4B). A total of 44 superfamilies were found to be common across the three non-whitefly genomes, of which 26 superfamilies were identified as DNA transposon superfamilies, four as LTRs, and 14 as LINEs. Unique superfamilies were also identified within each genome. DIPSY had the greatest number of unique superfamilies at 15 (three DNA transposon superfamilies, one LTR superfamily, five LINE superfamilies, and six SINE superfamilies) while MYPER had the least at five (one DNA transposon superfamily, three LTR superfamilies, and one LINE superfamily).

A further comparison of the superfamilies was performed between the 63 common superfamilies found within the whitefly genomes and the three non-whitefly hemipteran genomes analyzed by the same workflow methodology (Figure 4C). A total of 35 superfamilies were identified as common across all the groups (21 DNA transposons, four LTRs, and ten LINEs). Nine

superfamilies were identified to be present in the three non-whitefly genomes which were not identified in the superfamilies common to all the whitefly genomes. However, seven of these nine superfamilies were found to be present in one or two of the three whitefly genomes analyzed. Lastly, a total of 11 superfamilies of the 63 superfamilies common to all whitefly genomes were uniquely identified in them and not found in any of the three other hemipteran genomes. Nine of these 11 superfamilies represent DNA transposons (CMC-Chapaev-3, EnSpm/CACTA, ISL2EU, Kolobok, Mariner, PIF-Spy, Sola-2, TcMar-Tc4, and Zator) while the remaining two were LINE superfamilies (Nimb and L2B).

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 members of the species complex were the focus of this analysis (MEAM1, MED/Q, and SSA-ECA). Our results indicate that some previously published data contains erroneous identifications of TE distribution. An improved and standardized TE annotation workflow will allow a more accurate analysis of the distribution of TE across the whitefly species complex.

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 the 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 [58], yet 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 result from this study shows a 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 size

A positive correlation between TE content and genome size has previously been reported in arthropod genomes [34, 59, 60] as well as other genomes [18, 44, 61]. An arthropod wide study conducted by Petersen et al. [34] was the most extensive showing the association of genome 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 coverage) whilst the smallest genome studied (*Belgica antarctica* 89.54 Mbp) have the lowest TE proportion (2.58% genome coverage).

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 (Figure 2). MYPER, 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, 59, 62], 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 were linked to genome expansion; however, the exact order(s) of TEs which contributed to the expansion remains unclear [60]. An association of a specific TE orders (DNA transposons) and genome size was identified in the *Clitarchus hookeri* genome [59]; 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, 59, 60]. DNA transposons can act as cis-regulatory elements which increases 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. DNA transposons were identified to be the most abundant TE in the *C. hookeri* genome and comparison against the three other polyneoptera genome shows an association of DNA transposons and genome size [59]. 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 have been reported in MEAM1 [14] and SSA-ECA [15] genomes but was never explored. The presence of common and unique DNA transposon superfamilies across the whitefly genomes highlights the importance of the 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, 68]. The progenies of these insects became sterile and had an increase in the frequency of mutations and chromosome rearrangement [28, 68]. LINEs have been observed to successfully maintain themselves thru their host organism's evolutionary lifetime [65, 66, 69]. Site-specific insertion of R1 and R2 LINE superfamilies near the 28S ribosomal RNA (rRNA) genes ensured its propagation while there is also evidence of another LINE superfamily successfully maintaining itself through non-site-specific insertion [65–67, 69]. 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, 60, 70]. The

consequences of the low distribution of LINEs within the whitefly species complex are unknown and the 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, 71] 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 MYPER 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 the seven of the 73 genomes included in the study. It is possible that there are 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. reported 2.07% genome coverage of SINEs within the *Heliconius melpomene* genome and 9.41% within the *B. mori* genome. In the *H. melpomene* TE analysis, Lavoie et al. [70] identified more at 8.22% genome coverage; while in the *B. mori* TE analysis done by Osanai-Futahashi et al. [64], 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 warrant further investigation and validation to enable improved classification of these elements. Since 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 still remain the most abundant as more than half of the identified elements in the species complex are DNA transposons.

The 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. 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 remains unknown. These issues highlight the need to further explore 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 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; *Acyrtosiphon pisum* [ACPIS] [77], *Diaphorina citri* [DIPSY] [78], and *Myzus persicae* [MYPER] [79].

The MEAM1 (*B. tabaci* Middle East-Asia Minor 1) draft genome assembly was obtained from GenBank under the accession number GCA_001854935.1. DNA used for this MEAM1 assembly was extracted from 6500 male (haploid) whiteflies originating from an isofemale colony, i.e. generated from a single female [14]. Extracted DNA was sequenced using Illumina and PacBio technologies. Three Illumina paired-end reads were produced using the Illumina HiSeq 2500 system, and one PacBio library was produced using the PacBio RSII Sequencing system. The assembly was a *de novo* assembly with 300x coverage and using Platanus v1.2.1 assembly method. The libraries from the paired-end reads were assembled first using Platanus and gaps were further filled in using the PBJelly software with the PacBio long reads [14].

The MED/Q (*B. tabaci* Mediterranean) was obtained from www.gigadb.org/dataset/100286. DNA had been extracted from ~5000 mixed adult (haploid males and diploid females) whiteflies from Beijing, China [16]. The DNA sample was then used to construct Illumina paired-end libraries which were sequenced using Illumina HiSeq2000 using 100 bp paired-end reads. Low-quality reads were filtered out, before assembly using SOAPdenovo. Thirteen BAC (Bacterial Artificial Chromosome) libraries were produced, and each was used to assemble whole genome shotgun reads that had their gaps filled with the paired-end and mate-pair libraries. The draft assembly was produced from the overlapping sequences of the genome shotgun reads and BAC reads [16].

The SSA1-ECA genome is the most recently released *B. tabaci* genome included (<ftp://www.whiteflygenomics.org/pub/whitefly/SSA-ECA/v1.0/>) [15]. DNA was extracted from ~1,050 adult male whiteflies that were collected from cassava plants in a single field in Tanzania. Paired-end and mate-pair libraries were sequenced using the Illumina HiSeq2500 system. The assembly performed was a *de novo* assembly with Platanus. The assembly was then further improved by identifying contaminants using BLAST, and Pilon to identify base errors, mis-assemblies, and to fill in the gaps [15].

The *Acyrtosiphon pisum* [ACPIS] genome was obtained from NCBI under the project accession ABLF01000000. The aphids used to assemble the genome came from a single generation of inbreeding which was called the LSR1.AC.G1 aphid line. The DNA sample was taken from an LSR1.AC.G1 aphid line that was grown from a single female and was treated with ampicillin to remove the symbiont *Reigella insecticola* [77]. Sanger sequencing was used to produce the reads and was assembled using the Atlas assembly pipeline with a 6.2x coverage [77].

The *Diaphorina citri* [DIPSY] genome was obtained from NCBI under the project code AWGM01000000. The version Diaci1.1 genome build was used for this study. DNA had been extracted from a *D. citri* collected from a citrus grove in Ft. Pierce Florida [78]. The genome was sequenced using a combination of Illumina HiSeq and supplemented with PacBio technology and was assembled using a PBJelly v2013 pipeline.

The *Myzus persicae* [MYPER] genome was also obtained from NCBI (project code LXJY01000000). The Clone G006 lineage was used in the construction of the library used for the genome assembly. The G006 lineage was maintained on *Brassica oleracea* var. Wisconsin golden acre [79]. Tissue was extracted from 81 females and 134 males, and was used to produce a single paired-end library and two mate-pair libraries [79]. The genome was sequenced using Illumina HiSeq with 200.0x coverage and was assembled using ALLPATHS-LG2 v.r40324.

Repeat Identification

The workflow performed in this study for creating a species-specific repeat library. The genome to be studied is first submitted to MITTracker (<https://github.com/INTABiotechMJ/MITE-Tracker>) [80] and TransposonPSI_08222010 (<http://transposonpsi.sourceforge.net/>) for the initial step of the identification. The result from MITTracker is then masked

from the genome and submitted for LTR identification. Results from MITETracker and LTR identification are gathered and masked from the genome before submitting to RepeatModeler v1.0.11 [81]. This 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 genomertools v1.5.9 (LTRHarvest and LTRDigest), RepeatModeler v1.0.11, and RepeatMasker v4.1.1 [82]. 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 [83, 84]. The clustering algorithm by USEARCH is implemented to identify clusters of sequences that are within $\geq 80\%$ identity of each other. A consensus sequence is then produced from each of the clusters to obtain a representative sequence. 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/rmhuley/RepeatModeler/blob/master/RepeatClassifier>) and the unclassified sequences were subsequently submitted to the web browser version of Censor [85]. 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 [86] and PASTECClassifier v1.0 [87]. Manual curation was done to analyse the result of both tools and classification which had the consensus of both tools were accepted as the final classification. Results from the homology-based classification, the consensus classification of TEClass and PASTECClassifier, 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 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 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

ACPIS: *Acyrtosiphon pisum*

DIPSY: *Diaphorina citri*

DNA: Deoxyribonucleic Acid

GTF: Gene transfer format

LINE: Long interspersed nuclear elements

LTR: Long terminal repeat

Mbp: Mega base pairs

MITE: Miniature inverted repeat transposable element

MYPER: *Myzus persicae*

NonLTR: Non-long terminal repeat

RNA: Ribonucleic Acid

rRNA: Ribosomal RNA

SINE: Short interspersed nuclear elements

TE: Transposable element

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

MEAM1, ACPIS, DIPSY, and MYPER 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 and associated downstream analysis script are available from the corresponding author upon request.

Competing interests

All authors have no competing interest to declare

Funding

The research was made possible through the University of Greenwich Vice-Chancellor Scholarship.

Authors' contributions

JPAS gathered, analysed, and interpreted the data used in this study. SOS assisted in the initial preparation of the TE identification workflow. JPAS developed and tested the workflow. PMV, SB, and SES verified the methods used in the study and reviewed the manuscript. PMV provided Bioinformatics advice and support. SES supervised JPAS in the investigation of the topic, provided technical expertise and supervised the findings of this work. JPAS wrote the main manuscript. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

The authors would like to thank the University of Greenwich and the Natural Resources Institute.

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. doi:10.1080/07352680500365257.
2. Naranjo SE, Chu C-C, 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. doi: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.
4. Martin JH, Mound LA. An annotated check list of the world's whiteflies (Insecta: Hemiptera: Aleyrodidae). 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-Olive E, Sanchez-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. doi:DOL: 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* (2004). 2021. doi: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. doi:10.1111/mec.14865.
10. Malka O, Feldmesser E, van Brunschot S, Santos-Garcia D, Han W-H, 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. doi:https://doi.org/10.1111/eva.13162.
11. 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. Chi Y, Pan L-L, Bouvaine S, Fan Y-Y, Liu Y-Q, Liu S-S, 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. doi:https://doi.org/10.1016/j.virol.2019.11.013.
13. Fan Y-Y, Zhong Y-W, Zhao J, Chi Y, Bouvaine S, Liu S-S, et al. *Bemisia tabaci* Vesicle-Associated Membrane Protein 2 Interacts with Begomoviruses and Plays a Role in Virus Acquisition. *Cells*. 2021;10.
14. Chen W, Hasegawa DK, Kaur N, Klot 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. doi: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. doi: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. doi: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:evab062. doi:10.1093/gbe/evab062.

18. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*. 2002;115:49–63. doi: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. doi: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. doi: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. doi:10.1186/s13059-015-0729-7.
22. Finnegan DJ. Eukaryotic transposable elements and genome evolution. *Trends Genet*. 1989;5:103–7.
23. Finnegan DJ. Transposable elements. *Curr Opin Genet Dev*. 1992;2:861–7.
24. Jurka J, Kapitonov V V, 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. *Molecular Phylogenetics and Evolution*. 2015;86:90–109.
27. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet Syst*. 2018.
28. Galun E. Transposable Elements. Dordrecht: Springer Netherlands; 2003. doi: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. doi:10.1146/annurev.genet.40.110405.090448.
30. Eickbush TH. Retrotransposons. In: Brenner S, Miller JHBT-E of G, editors. New York: Academic Press; 2001. p. 1699–701. doi: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. doi:https://doi.org/10.1016/j.cub.2012.04.025.
33. Kazazian Jr HH, Scott AF. “Copy and paste” transposable elements in the human genome. *J Clin Invest*. 1993;91:1859–60. doi:10.1172/JCI116400.
34. Petersen M, Armisen 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. doi:10.1186/s12862-018-1324-9.
35. Parisot N, Vargas-Chavez C, Goubert C, Baa-Puyoulet P, Balmand S, Beranger L, et al. The genome of the cereal pest Sitophilus oryzae; a transposable element haven. *bioRxiv*. 2021;:2021.03.03.408021. doi:10.1101/2021.03.03.408021.
36. Griffiths AJ, Lewontin R, Gelbart WM, Miller J. Modern Genetic Analysis. Second. W.H. Freeman & Co. Ltd; 2002. https://www.ncbi.nlm.nih.gov/books/NBK21254/.
37. Gilbert C, Peccoud J, Cordaux R. Transposable Elements and the Evolution of Insects. *Annu Rev Entomol*. 2021;66:355–72. doi: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. doi: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. doi: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. doi: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 M-J, Zhou Q-Z, Zhang H-H, Tong X, Lu C, Zhang Z, et al. iMITEdb: the genome-wide landscape of miniature inverted-repeat transposable elements in insects. Database (Oxford). 2016;2016:baw148. doi: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. doi: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. Genome Biol. 2018;19:199. doi: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. doi: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 LP – 1093. doi: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. doi: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. doi:10.1101/gr.185579.114.
51. Goubert C, Modolo L, Vieira C, ValienteMoro C, 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. doi:10.1093/gbe/evv050.
52. Hill T. Transposable element dynamics are consistent across the &Drosophila& phylogeny, despite drastically differing content. bioRxiv. 2019;651059. doi:10.1101/651059.
53. Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable elements in *Drosophila*. Mob DNA. 2020;11:23. doi: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. doi: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 Y-W, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci. 2011;108:7884 LP – 7889. doi:10.1073/pnas.1104208108.
58. Bregliano JC, Picard G, Bucheton A, Pelisson A, Lavigne JM, L&Heritier P. Hybrid dysgenesis in *Drosophila melanogaster*. Science (80-). 1980;207:606 LP – 611. doi:10.1126/science.6766221.
59. 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. doi:10.1186/s12864-017-4245-x.
60. Talla V, Suh A, Kalsoom F, Dincă 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. doi:10.1093/gbe/evx163.
61. Naville M, Henriët S, Warren I, Sumic S, Reeve M, Volff J-N, et al. Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. Curr Biol. 2019;29:1161-1168.e6.

62. 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. doi:10.1098/rsbl.2016.0407.
63. Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, et al. High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol.* 2019;107:53–62. doi:https://doi.org/10.1016/j.ibmb.2019.02.002.
64. 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. doi:https://doi.org/10.1016/j.ibmb.2008.05.012.
65. Lathe WC 3rd, 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 3rd, 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. doi:10.1093/oxfordjournals.molbev.a025732.
67. Jakubczak JL, Burke WD, Eickbush TH. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc Natl Acad Sci.* 1991;88:3295 LP – 3299. doi:10.1073/pnas.88.8.3295.
68. 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.
69. 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. doi:10.1186/1471-2148-7-112.
70. 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. doi:10.1186/1759-8753-4-21.
71. Ohshima K, Okada N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res.* 2005;110:475–90. doi:10.1159/000084981.
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. doi: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.* 2020;117:9451 LP – 9457. doi:10.1073/pnas.1921046117.
74. Peccoud J, Loiseau V, Cordaux R, Gilbert C. Massive horizontal transfer of transposable elements in insects. *Proc Natl Acad Sci.* 2017;114:4721 LP – 4726. doi:10.1073/pnas.1621178114.
75. Martin JH, Mifsud D, Rapisarda C. The whiteflies (Hemiptera: Aleyrodidae) of Europe and the Mediterranean Basin. *Bull Entomol Res.* 2000;90:407–48.
76. Xia J, Guo Z, Yang Z, Han H, Wang S, Xu H, et al. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell.* 2021;184:1693-1705.e17.
77. Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D, et al. Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol.* 2010;8.
78. 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 (Oxford).* 2017;2017:bax032. doi:10.1093/database/bax032.
79. 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. doi:10.1186/s13059-016-1145-3.
80. 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. doi:10.1186/s12859-018-2376-y.
81. Smit A, Hubley R. RepeatModeler Open-1.0. 2008. <http://www.repeatmasker.org>.
82. Campbell MS, Law M, 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. doi:10.1104/pp.113.230144.

83. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26:2460–1. doi:10.1093/bioinformatics/btq461.

84. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods*. 2013;10:996. <https://doi.org/10.1038/nmeth.2604>.

85. 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. doi:10.1186/1471-2105-7-474.

86. Abrusán G, Grundmann N, Demester L, Makalowski W. TEclass - A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*. 2009;25:1329–30.

87. 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>.

Figures

Figure 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.

Figure 2

Percent coverage of transposable elements and genome size

Each genome was plotted in relation to their TE coverage and genome size. TE coverage 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 ($r = 0.93$)

Figure 3

Percent coverage of each order of transposable elements in the non-whitefly and whitefly group

Box plots comparing percent genome coverage 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; DIPSY having the highest at 3% while none was detected in MYPER.

Figure 4

Frequency of repeat families identified across the genomes

(A) The frequency of repeat superfamilies identified in the whitefly genomes. A total of 63 superfamilies (39 DNA transposons, six LTRs, 16 LINEs, and two SINEs) are found to be common across the three whitefly genomes. **(B)** The frequency of repeat superfamilies identified in the non-whitefly genomes and found in the other non-whitefly genomes. A total of 44 superfamilies are found to be common across the three non-whitefly genomes: 26 DNA transposons, four LTRs, and 14 LINEs. **(C)** The frequency of the 63 common repeat superfamilies found across the three whitefly genomes and found across the three non-whitefly genomes. A total of 11 superfamilies were uniquely identified in the whitefly genomes while 35 superfamilies are found present in all the genomes.