Diversity of short interspersed nuclear elements (SINEs) in lepidopteran insects and evidence of horizontal SINE transfer between baculovirus and lepidopteran hosts

Guangjie Han1,2, Nan Zhang1, Heng Jiang1, Xiangkun Meng1, Kun Qian1, Yang Zheng1, Jian Xu2* and Jianjun Wang1,3*

Abstract

Background: Short interspersed nuclear elements (SINEs) belong to non-long terminal repeat (non-LTR) retrotransposons, which can mobilize dependent on the help of counterpart long interspersed nuclear elements (LINEs). Although 234 SINEs have been identified so far, only 23 are from insect species (SINEbase: http://sines.eimb.ru/).

Results: Here, five SINEs were identified from the genome of Plutella xylostella, among which PxSE1, PxSE2 and PxSE3 were tRNA-derived SINEs, PxSE4 and PxSE5 were 5S RNA-derived SINEs. A total of 18 related SINEs were further identified in 13 lepidopteran insects and a baculovirus. The 3′-tail of PxSE5 shares highly identity with that of LINE retrotransposon, PxLINE1. The analysis of relative age distribution profiles revealed that PxSE1 is a relatively young retrotransposon in the genome of P. xylostella and was generated by recent explosive amplification. Integration pattern analysis showed that SINEs in P. xylostella prefer to insert into or accumulate in introns and regions 5 kb downstream of genes. In particular, the PxSE1-like element, SINPVYSE1, in Spodoptera litura nucleopolyhedrovirus II genome is highly identical to SfSE1 in Spodoptera frugiperda, SlittSE1 in Spodoptera littoralis, and SlituSE1 in Spodoptera litura, suggesting the occurrence of horizontal transfer.

Conclusions: Lepidopteran insect genomes harbor a diversity of SINEs. The retrotransposition activity and copy number of these SINEs varies considerably between host lineages and SINE lineages. Host-parasite interactions facilitate the horizontal transfer of SINE between baculovirus and its lepidopteran hosts.

Keywords: Short interspersed nuclear element (SINE), Horizontal transfer, Plutella xylostella, Retrotransposon, Long interspersed nuclear elements (LINEs)
Background
Short interspersed nuclear elements (SINEs) are Class I transposable elements (TEs) that propagate by a copy-and-paste mechanism [1, 2]. SINEs are evolutionarily derived from endogenous RNA polymerase III (Pol III) transcripts [3]. While mammalian SINEs, such as B1 and Alu, are originated from 7SL RNAs, other eukaryotes primarily harbor tRNA-like SINEs [4], and SINEs originated from 5S rRNA have been found in zebrafish, fruit bats, and springhare [5, 6]. Recently, SINEs derived from small nuclear RNA (snRNA) (SINEU) and the 3′-end of the large ribosomal subunit (LSU or 28S rDNA) (SINE28) have been identified in avian, crocodilian and mammalian genomes, respectively [7–9]. The characteristic features of SINEs include a 5′ terminal RNA-related region which contains an internal Pol III promoter, a central region, and a 3′-tail that is of variable length and recognized by the reverse transcriptase (RT) of autonomous partner long nuclear interspersed element (LINE) during retrotransposition [3]. The SINE promoters originated from tRNA and 7SL RNA comprise box A and B motif, whereas 5S rRNA-derived SINE promoters have three boxes such as A, IE and C [10].

As non-autonomous retrotransposon, the replication rate and survival of a SINE is dependent on the partner LINE activity, and its genomic copy number varies greatly between families and host species. For example, as high as 1 million copies of Alu elements have been identified in the human genome [11], whereas only two copies of ZmSINE3 were detected in Zea mays [12]. On the other hand, the number of SINE families within a genome is also highly variable, ranging from a single SINE family in the Vitaceae to 22 SINE families in the Amaranthaceae [13]. Interestingly, unlike LINEs, the distribution of a SINE family is generally restricted to a certain taxonomic group such as orders/families [3, 4, 14], suggesting that SINEs are one of the major genetic elements that determine a clade-specific genomic composition.

Transposable elements play an important role in the epigenetic regulation of the genome and generation of genomic novelty. A growing body of evidence has recently accumulated indicating that SINEs have a deep impact on genome organization and gene structure by generating regulatory elements for gene expression [15, 16], exon skipping and alternative splicing [17], alternative polyadenylation signals [18, 19], and even functional RNA genes [20, 21]. For example, an Alu SINE inserted into human pluripotency-associated transcript 5 (HPAT5) regulated related microRNAs through its let-7 binding site, which is essential for inner cell mass formation during early embryonic development [22].

While SINEs have been well characterized in human [23], other mammals [24] and plants [25], and currently about 200 SINE families/subfamilies are identified in various clades in Metazoa, as reported in Repbase [26] and in SINEBase [2], information on insect SINEs is still limited [27–33]. Recent improvements in both genome sequencing and assembly methodologies have led to increasing high-quality insect genome assemblies, which provides the opportunity to identify novel SINEs. However, due to their minimal sequence feature, the lack of coding capacity, and high sequence heterogeneity, annotations of SINEs are often incomplete or missing. Here, we described three tRNA-derived SINE families and two 5S rRNA-derived SINE families in the diamondback moth (DBM), Plutella xylostella (L.), which is one of the most damaging insect pests of cruciferous vegetables around the world. We investigated the structures and insertion regions of these SINEs. The distribution of these SINEs in other lepidopteran insect species was also surveyed.

Results
Novel tRNA-derived SINE retrotransposons, PxSE1, PxSE2 and PxSE3, in P. xylostella
A novel tRNA-derived SINE, PxSE1, was identified by homology search in DBM genome database (Additional file 1: Figure S1). A total of 68 full length copies homologous to PxSE1 were used to reconstruct the consensus sequence of PxSE1 (Accession numbers: MW068006-MW068073). The PxSE1 is 263 bp long, includes GT dinucleotide repeats at 3′-tail and a 72-bp tRNA-related region at the 5′-end with 64% identity to 72-bp tRNA<sup>Arg</sup> of Drosophila melanogaster, which contains box A and box B of the RNA Pol III promoter (Figs. 1 and 2a, Additional file 2: Figure S2). The boundary of PxSE1 was further defined by the alignment of a PxSE1 element and its empty site sequence (Additional file 3: Figure S3). Using the PxSE1 as the query, a total of 6208 copies were identified in DBM genome (Table 1). The average divergence is 0.035 in all PxSE1 copies (Table 1), indicating a recent invasion time.

Using PxSE1 as query, two additional tRNA-derived SINEs, PxSE2 and PxSE3, were identified by database searches. The consensus sequences of PxSE2 and PxSE3 were reconstructed using the same methods as described above (Accession numbers: MW068074-MW068156, Additional file 2: Figure S2 and Additional file 3: Figure S3). The PxSE2 is 263 bp long, includes a 143 bp 3′-end sequence, which is different from PxSE1, but has 67.5% identity with BmSE. The 72-bp tRNA-related region of PxSE2 is 93.4% identical to PxSE1 (Figs. 1 and 2a). Interestingly, PxSE2 has a 44 bp conserved central domain with 93.2% identity to PxSE1 (Fig. 2a). The PxSE3 is 339 bp long, includes a 72-bp tRNA-related region with 66.5% identity to tRNA<sup>Arg</sup> of D. melanogaster, and has 79.3% identity with the 228 bp sequence at 5′-end of
SINE2-1_Pxu from *Papilio xuthus* [34] (Figs. 1 and 2b). The copy numbers of PxSE2 and PxSE3 were 5056 and 5158 in DBM genome, respectively (Table 1). The average divergence of PxSE2 and PxSE3 were 0.071 and 0.089, respectively (Table 1).

Distribution of PxSE1, PxSE2 and PxSE3 in other species

BLAST searches were performed to detect PxSE1, PxSE2 and PxSE3 sequences in insect species other than *P. xylostella*. In total, homologous sequences of PxSE1, PxSE2 and PxSE3 were identified in five, two and seven Lepidoptera insects, respectively (Accession numbers: MW068230-MW069451, Additional file 2: Figure S2), among which MsSE2 in *Manduca sexta* showed the highest copy numbers of 16,157, whereas only 533 copies of CsSE1 were detected in the genome of *Chilo suppressalis* (Table 1). The consensus sequences of these elements vary in size from 252 bp to 333 bp and have different 3’-tails. Differently, the consensus sequences of EpSE1 did not contain poly(A), poly(T) or simple sequence repeats at 3’-end. The average divergence varied from 0.035 to 0.13 (Table 1). Although PxSE2- and PxSE3-like elements were not identified in non-insect species, a PxSE1-like element, SINPVSE1, was detected in *Spodoptera litura* nucleopolyhedrovirus II (EU780426.1: 30485–30735), which was located within ORF27 encoding an unknown protein.

Multiple sequence alignment of the consensus sequences showed that the evolutionary divergence varied from 0.003 to 0.436. The highest identity (99.7%) was observed between PmSE1 in *Papilio machaon* and PxSE1 in *Papilio zelicaon*, whereas MsSE1 in *M. sexta* and CfSE1 in *Choristoneura fumiferana* showed the highest evolutionary divergence (0.436) (Additional file 4: Figure S4).

Two 5S rRNA-derived SINEs, PxSE4 and PxSE5, in *P. xylostella* and related species

Using HaSE3 as a query [33], BLAST searches revealed two 5S rRNA-derived SINEs, PxSE4 and PxSE5, in DBM (Accession numbers: MW068157-MW068229, Figs. 1 and 2c). The boundary of PxSE4 and PxSE5 was further defined by the alignment of single PxSE element and its empty site sequence (Additional file 2: Figure S2 and Figure S3). PxSE4 and PxSE5 are both 389 bp in length and shared high identity of 250 bp sequence at 5’-end but are different at the 3’-end. The promoter regions of PxSE4 and PxSE5 include the specific A box, IE and C boxes, and shared about 63% identity with 5S rRNA of *Bombyx mori*, indicating that they are 5S rRNA-derived SINEs (Fig. 2c). The copy numbers and average divergence of PxSE4 and PxSE5 were 4415 and 1952, 0.078 and 0.132, respectively (Table 1).

Interestingly, we found a LINE element PxLINE1.1 (NW_011952036.1: 552486–555713) with its 43-bp 3’-end being 84% identical to that of PxSE5 (Fig. 2). Thus, this region was designated as 3’-LINE-related region (Fig. 1). The PxLINE1.1 element was 3228 bp long, flanked by 13 bp target site duplications (TSDs), encoded L1_EN (Endonuclease domain of the non-LTR retrotransposon LINE-1) and RT domain, and was terminated by ATGT tetranucleotide repeats in the short 3’ untranslated region (3’ UTR) (Fig. 3). Additional eight copies were found to be 96.1 to 99.7% identical to PxLINE1.1 in *P. xylostella*. Specifically, one copy (AHIO01028576.1:13049_14357) from WGS was inserted as a 1686 bp fragment, which shared 71.8% identity with mariner-8_BM from *B. mori* [35] (Table 2 and Additional file 5: Figure S5). Sequences sharing 63 to 82% identity with the 1580 bp fragments at the 3’-end of PxLINE1.1 were also found in the other 7 lepidopteran insect genomes (Additional file 6: Figure S6).
Fig. 2 (See legend on next page.)
The *PxSE4* and *PxSE5* sequences were used as queries to search against the whole genome shotgun (WGS) and expressed sequence tags (EST) database using BLASTN. Three elements, *LaSE2*, *CsSE2* and *ObSE2*, with high identities to *PxSE4* were found in genomes of *Lerema accius*, *C. suppressalis* and *Operophtera brumata*, respectively (Accession numbers: MW068230-MW069451, Additional file 2: Figure S2). In particular, the 115-bp fragment at 5′-end of *ObSE2* is different from *PxSE4*, whereas the central 122-bp fragment shares highly identity with *PxSE4* (Additional file 4: Figure S4G and Additional file 7: Figure S7B). The 75-bp fragment at 5′-end of *ObSE2* is 54.2% identical to the 72-bp tRNA of *D. melanogaster*, but different from *PxSE1* (Additional file 7: Figure S7A). However, no simple repeat sequences were found at the 3′-ends of the *ObSE2*. While we did not find *PxSE5*-like elements in other insects, the 56-bp fragment at 3′-end of *PxSE5* and *SfSE1* shared 89.6% identity (Additional file 4: Figure S4I).

**Transpositional burst of SINEs**

Due to the accumulation of random mutations over time, evolutionarily ancient SINE families have a lower sequence identity among copies, whereas SINE families with recent or ongoing transposition harbor relatively homogeneous copies [12]. To evaluate the periods of transpositional activity and relative age of SINE copies per family of SINEs, we performed a pairwise

### Table 1: Novel SINE elements identified in this study

| SINE Family | Species                  | RNA Origin | Consensus Length | Tail  | Copy Number* | Divergence | Resource |
|-------------|--------------------------|------------|------------------|-------|--------------|------------|----------|
| PxSE1       | *Plutella xylostella*/*Yponomeutoidea* | tRNA       | 263              | (GT)n | 6208 (68)    | 0.035      | WGS      |
| MsSE1       | *Manduca sexta*/*Bombycoidea*    | tRNA       | 267              | (GT)n | 7513 (133)   | 0.091      | WGS      |
| SfSE1       | *Spodoptera frugiperda*/*Noctuoidea* | tRNA       | 298              | (ATGT)n | 11,117 (79) | 0.130      | WGS      |
| SfNPYSE1    | *Spodoptera litura* nucleopolyhedrovirus II | tRNA       | 260              | (TGTAT)n | 1 (1)       | ND         | Nr/nt    |
| SlituSE1    | *Spodoptera litura*/*Noctuoidea* | tRNA       | 259              | (ATGGT)n | ND(N)      | ND         | EST      |
| SlittSE1    | *Spodoptera littoralis*/*Noctuoidea* | tRNA       | 270              | (ATGGT)n | ND(N)      | ND         | EST      |
| CFSE1       | *Choristoneura fumiferana*/*Tortricoidea* | tRNA       | 252              | (ATT)n  | ND(N)      | ND         | EST      |
| PxSE2       | *P. xylostella*            | tRNA       | 263              | (ATT)n  | 5056 (33)   | 0.071      | WGS      |
| ObSE1       | *Operophtera brumata*/*Geometroidea* | tRNA       | 275              | (TATT)n | 4521 (120)  | 0.066      | WGS      |
| CsSE1       | *Chilo suppressalis*/*Pyraloidea* | tRNA       | 287              | (TATT)n | 533 (125)   | 0.036      | WGS      |
| PxSE3       | *P. xylostella*            | tRNA       | 339              | (GAAT)n | 5158 (50)   | 0.089      | WGS      |
| MsSE2       | *M. sexta*                | tRNA       | 333              | (TAT)n  | 16,157 (126) | 0.090      | WGS      |
| PgSE1       | *Papilio glaucus*/*Papilionoidea* | tRNA       | 303              | (TAT)n  | 1751 (66)   | 0.097      | WGS      |
| PmSE1       | *Papilio machaon*/*Papilionoidea* | tRNA       | 310              | (GAT)n  | 5740 (138)  | 0.098      | WGS      |
| LaSE1       | *Lepera accius*/*Hesperioidea* | tRNA       | 319              | (GAT)n  | 3832 (52)   | 0.107      | WGS      |
| PzSE1       | *Papilio zelicaon*/*Papilionoidea* | tRNA       | 293              | (GAT)n  | ND(N)      | ND         | TSA      |
| EpSE1       | *Erynnis propertius*/*Hesperioidea* | tRNA       | 300              | ND      | ND(N)      | ND         | TSA      |
| SeSE1       | *Spodoptera exigua*/*Noctuidae* | tRNA       | 291              | (GAT)n  | ND(N)      | ND         | TSA      |
| PxSE4       | *P. xylostella*            | SS rRNA    | 389              | (TGTA)n | 4415 (50)   | 0.078      | WGS      |
| LaSE2       | *L. accius*                | SS rRNA    | 348              | ND      | 1214 (45)   | 0.101      | WGS      |
| CsSE2       | *C. suppressalis*          | SS rRNA    | 294              | (TGTA)n | 532 (169)   | 0.021      | WGS      |
| ObSE2       | *O. brumata*              | tRNA       | 255              | (CGAA)n | 863 (65)    | 0.012      | WGS      |
| PxSE5       | *P. xylostella*            | SS rRNA    | 389              | (ATGT)n | 1952 (23)   | 0.132      | WGS      |

*ND* not determined

*the number in bracket is the number of copies used to reconstruct the consensus sequences*
Fig. 3 The nucleotide sequence and conceptual translation of the partner LINE element, PxLINE1.1, for PxSE5. Flanking direct repeats are indicated in lowercase. The nucleotides of TSD are indicated with the wavy line. The nucleotides of 3′ tail sequence are indicated with the straight line.

Han et al. BMC Genomics (2021) 22:226

Page 6 of 16
comparison of SINE copies with the consensus sequences of respective family and grouped them into intervals from 80 to 100% identity. As shown in Fig. 4, 4796 of 6208 copies of *PxSE1* show more than 95% identity to the consensus sequence, of which 223 copies are 100% identical to *PxSE1* consensus sequence (Additional file 8: Table S1), indicating a recent transpositional burst. A strong transposition peak with high identity values is also observed in *PxSE2*, *PxSE3*, and *PxSE4*. However, *PxSE5* shows high numbers of diverged copies, and only 49 copies (2.5%) of *PxSE5* have more than 95% identity with its consensus sequence (Fig. 4).

The activity profiles deduced from similarity intervals of SINEs in other lepidopteran species revealed a recent transpositional burst of *CsSE1* and *CsSE2* in *C. suppressalis* and *ObSE2* in *O. brumata*, whereas *LaSE1* and *SjSE1* harbour diverged copies and only few young ones (Fig. 4 and Additional file 9: Figure S8). High number of copies with a wide range of identity values were observed in *MsSE1*, *MsSE2*, *PgSE1*, *PmSE1* and *LaSE2* (Fig. 4 and Additional file 9: Figure S8). Due to few copies in related EST and transcriptome shotgun assembly (TSA) databases, the distribution profiles of copy identity in *SlitSE1*, *SlittSE1*, *CjSE1*, *PzSE1*, *EpSE1* and *SeSE1* were not subject to analysis.

**Table 2** Copies with high identity to *PxLINE1.1* in *P. xylostella*

| Subject No.         | Identity % | Length | Start  | End    | Evalue | Bit score |
|---------------------|------------|--------|--------|--------|--------|-----------|
| AHIO01036046.1      | 99.6       | 3203   | 31,532 | 34,758 | 0      | 5838      |
| AHIO01028688.1      | 99.5       | 3203   | 1034   | 4236   | 0      | 5819      |
| AHIO01028673.1      | 99.4       | 3203   | 10,866 | 7665   | 0      | 5810      |
| AHIO01016682.1      | 99.3       | 3203   | 40,175 | 47,520 | 0      | 5243      |
| AHIO01031207.1      | 99.2       | 3207   | 51     | 3252   | 0      | 5749      |
| AHIO01033561.1      | 99.1       | 3216   | 43,696 | 40,488 | 0      | 5243      |
| AHIO01003557.1      | 99.7       | 4408   | 11,175 | 15,783 | 0      | 5067      |

Contribution of SINEs to gene and genome evolution in *P. xylostella*

The integration pattern relative to the annotated genes in the genome of *P. xylostella* was analyzed. A total of

![Fig. 4 Examples for the relative age distribution of SINE families in *P. xylostella*, *M. sexta*, *C. suppressalis*, *L. accius* and *O. brumata*. The abscissa showed the identities between each consensus sequence and the copies. The ordinate showed the copy numbers of sequence with the same identity. The same color represented the same family of SINE.](image)
2750 out of 6208 copies (44%) of \textit{PxSE1}, 2478 out of 5056 copies (49%) of \textit{PxSE2}, 2470 out of 5158 copies (48%) of \textit{PxSE3}, 2265 out of 4415 copies (51%) of \textit{PxSE4} and 902 out of 1952 copies (46%) of \textit{PxSE5} were found in introns (Fig. 5a). Similar proportions of the copies are distributed in regions 5kBp downstream of genes. Only two, five, eight and five copies of \textit{PxSE1}, \textit{PxSE2}, \textit{PxSE3}, \textit{PxSE4} and \textit{PxSE5} were found to insert into exonic regions, respectively (Fig. 5a). Among them, 11 copies are inserted into the coding regions (CDS), a copy is inserted into the 5′ UTR, and 13 copies are inserted into the 3′ UTR (Table 3). Most of these genes were annotated as enzymes or enzyme-associated proteins, and were related to signal transduction, splicing, metabolism. For example, a 261 bp copy \textit{PxSE2.2} of \textit{PxSE2} family from DBM genome (NW_011952011.1: 2273356–2273095) inserted into CDS of a gene encoding nitrogen permease regulator 3-like protein. The 21-bp fragment at 5′-end of \textit{PxSE2.2} contributed 7 amino acids to the N-terminus of the protein (Fig. 5b).

Further analysis revealed the insertion of copies of SINE families into introns of the same gene. As many as 60 elements inserted into introns of LOC105382892 gene, including 18, 14, 10, 11 and 7 copies of \textit{PxSE1}, \textit{PxSE2}, \textit{PxSE3}, \textit{PxSE4} and \textit{PxSE5}, respectively (Additional file 10: Figure S9). A total of 95 genes were found to be inserted with at least ten copies of SINE elements (Additional file 10: Figure S9D). Thus, the \textit{P. xylostella} SINE families contribute to structural variation in introns, which might influence the regulation of gene expression.

**Evolution and horizontal transposon transfer (HTT) of SINEs**

The phylogenetic tree of the 23 SINE consensus sequences showed that the SINEs with the same internal \textit{Pol III} promoter were clustered together, except \textit{ObSE2} SINE (Fig. 6a). Due to the high identity of \textit{PxSE1} and \textit{PxSE2} at 5′-ends, the clustering of related SINEs in different family, such as \textit{PxSE1}, \textit{PxSE2}, \textit{CfSE1}, \textit{ObSE1} and \textit{CsSE1}, is not surprising. The comparison of phylogenetic tree of \textit{PxSE3} family and the taxonomy tree of related host species [36, 37] (Fig. 6) suggests some degree of vertical transmission of \textit{PxSE3} family in lepidopteran insects. Interestingly, \textit{SINVPE1} and \textit{SfSE1} in \textit{Spodoptera frugiperda}, \textit{SlittSE1} in \textit{Spodoptera littoralis} and \textit{SlituSE1} in \textit{S. litura}, were clustered together (Table 1 and Fig. 6a). The orthologous outer flanking sequence of \textit{SINVPE1} were identified in \textit{Spodoptera eridania} nucleopolyhedrovirus isolate 251 and \textit{Spodoptera cosmioides} nucleopolyhedrovirus isolate VPN72, suggesting that SINVPE1 inserted into the genome of nucleopolyhedrovirus by HTT. In addition, the inter 5′-flanking sequence (about 800 bp) was found to share 95% identity
### Table 3: The annotation of SINEs copies integrated into CDS and untranslated regions (UTR) in *P. xylostella*

| Copies | Location | GeneID | Gene mapping | COG class annotation | Swissprot annotation | Nr annotation |
|--------|----------|--------|--------------|----------------------|---------------------|---------------|
| PxSE1.1| 436,420– 436,680 | LOC105383591 | CDS | Signal transduction mechanisms | Cyclic nucleotide-gated cation channel subunit A | uncharacterized protein |
| PxSE1.2| 46,181–46, 441 | LOC105394666 | CDS | Signal transduction mechanisms | Cyclic nucleotide-gated cation channel subunit A | uncharacterized protein |
| PxSE2.1| 800,671– 800,788 | LOC105384210 | 3’ UTR | ND | Transcription factor 25 homolog | transcription factor 25 |
| PxSE2.2| 2,273,356–2, 273,095 | LOC105380419 | CDS | ND | Nitrogen permease regulator 3-like protein | nitrogen permease regulator 3-like protein |
| PxSE2.3| 35,405–35, 544 | LOC105390425 | 3’ UTR | ND | Serine/threonine-protein kinase | serine/threonine-protein kinase grp-like |
| PxSE2.4| 250,511–250,756 | LOC105381765 | CDS | Replication, recombination and repair | DNA topoisomerase 3-alpha | uncharacterized protein |
| PxSE2.5| 266,277–266,424 | LOC105391817 | 3’ UTR | ND | Leucine-rich repeat serine/threonine-protein kinase 1 | uncharacterized protein |
| PxSE3.1| 274,387–274,287 | LOC105388973 | 3’ UTR | ND | Gamete and mating-type specific protein A | uncharacterized protein |
| PxSE3.2| 699,462–699,145 | LOC105386775 | CDS | ND | Uncharacterized protein | serine/arginine repetitive matrix protein 1-like |
| PxSE3.3| 490,234–490,329 | LOC105381296 | 3’ UTR | ND | Venom acid phosphatase Acph-1 (Precursor) | prostatic acid phosphatase |
| PxSE4.1| 1,260,998–1, 260,894 | LOC105380733 | 3’ UTR | General function prediction only | Protein suppressor of hairy wing | zinc finger protein 26-like |
| PxSE4.2| 63,533–63, 636 | LOC105391976 | 3’ UTR | General function prediction only | Mediator of RNA polymerase II transcription subunit 12 | uncharacterized protein |
| PxSE4.3| 31,764–31, 620 | LOC10539953 | 3’ UTR | General function prediction only | Ras-related protein Rab-24 | ras-related protein Rab-24-like |
| PxSE4.4| 787,511–787,311 | LOC105382052 | 3’ UTR | General function prediction only | WW domain-containing protein ZK1098.1 | transcription elongation regulator 1-like isoform X1 |
| PxSE4.5| 160,917–160,809 | LOC105388357 | CDS | ND | Serine/threonine kinase SAD-1 | PAS domain-containing serine/threonine-protein kinase |
| PxSE4.6| 1,160,410–1, 160,304 | LOC105398290 | 3’ UTR | General function prediction only | Ras-related protein Rab-24 | ras-related protein Rab-24-like |
| PxSE4.7| 819,669–819,771 | LOC105385258 | CDS | ND | Probable 4-coumarate-CoA ligase 2 | probable 4-coumarate-CoA ligase 3 |
| PxSE4.8| 217,956–218,125 | LOC105383715 | CDS | ND | Integrin alpha-PS3 light chain (Precursor) | integrin alpha-PS3-like |
| PxSE4.9| 306,942–307,066 | LOC105398401 | 5’ UTR | ND | ND | CLK4-associating serine/arginine rich protein-like |
| PxSE5.1| 101,368–101,480 | LOC105393324 | CDS | Coenzyme transport and metabolism | Molybdoenzyme synthase catalytic subunit | molybdoenzyme synthase catalytic subunit-like |
| PxSE5.2| 993,048–992,940 | LOC105383883 | CDS | ND | Salivary glue protein Sgs-3 (Precursor) | uncharacterized protein |
| PxSE5.3| 338,901–338,769 | LOC105387600 | 3’ UTR | ND | Alpha-(1,3)-fucosyltransferase C | alpha-(1,3)-fucosyltransferase C-like |
| PxSE5.4| 499,158–499,032 | LOC105388075 | 3’ UTR | ND | Cuticle collagen 1 (Precursor) | breast cancer metastasis-suppressor 1-like protein isoform X1 |

**ND** not determined
to the sequence (WNXL0100005.1: 248783–248238) of *Spodoptera exigua* genome (Additional file 11: Figure S10 and Fig. 7), putatively resulted from unknown horizontal gene transfer.

**Discussion**

The structure of three tRNA-derived and two 5S rRNA-derived SINE families

Up to now, more than 234 SINEs have been isolated from the genomes of human, mammals, reptiles, fishes, mollusks, fungus, green plants, and insects [2]. Based on current data, the tRNA-derived SINEs (~84%) were found widely in eukaryotic genomes [2]. Apart from the 5′ terminal head, SINEs also consist of typical body and variable repeated tail. In this study, we have identified three tRNA-derived SINE families, *PxSE1*, *PxSE2* and *PxSE3*. The 45 bp region in body region of *PxSE1* and *PxSE2* also showed high identity (93.3%) except the highly identical heads. Similarly, two 5S rRNA-derived SINEs, *PxSE4* and *PxSE5*, also shared 98.7% identity in 159 bp region of their bodies. Previous studies have found that the conserved bodies of SINE mainly include the V-domain, CORE-domain, Deu-domain, Nin-domain, Ceph-domain, Inv-domain, Pln-domain, Snail-domain, and Meta-domain [38–43]. However, the body regions identified in *PxSEs* are different from these known domains. A hypothesis has proposed that nonautonomous LINEs that have only 5′ and 3′ regions of original LINEs can be a source of enigmatic middle body of SINEs [1]. Hence, highly identical conserved central domains among different SINEs in the same species suggests that the conserved central domain may originated from the same LINE family and has been under strong selective constraint, which is important for reverse transcription. In addition, despite the high identity between *ObSE2* and 5S rRNA-derived *PxSE4*, *ObSE2* is a tRNA-derived SINE. Partner LINE SINEs can be composed of 5′ and 3′ regions of nonautonomous LINEs, and their 3′ tails will also exchange with other LINEs under the pressure of natural selection to facilitate rapid amplification [1]. The tail homologous to LINE is important for SINE, which allows the integration of new copies of SINE into the new genomic locations using the LINE RT [44]. LINE RT can specifically recognize the 3′ homologous SINE tails, indicated that SINE can be mobilized by the retrotransposition machinery of a partner LINE [45]. Here, nine novel LINE copies in *P. xylostella*, and seven LINEs in each of the lepidopteran insects were identified with 3′-end similar to that of *PxSE5* and *SfSE1*, suggesting that the LINE
Fig. 7 (See legend on next page)
identified in this study is an ancient retrotransposon and might widely exist in Lepidoptera insects. However, the 5’ regions among ShSE1, PxSE5 and PxLINE1 shared a large divergence, indicating these SINEs exploded after the exchange of 3’-tails. Moreover, the distinct 3’-end in other PxSEs suggests that these SINEs might be mobilized by other LINEs that were not identified yet.

Relative age and distribution of SINEs in Lepidoptera insect

The copy numbers of SINEs varies among different families and species. In P. xylostella, the copy numbers of SINEs of tRNA origin is relatively higher than that of 5S rRNA origin. In particular, the copy numbers of PxSE5 is only 1952. Previously, it was speculated that the type 1 promoter in 5S rRNA is more dependent on upstream signals than the type 2 promoter in tRNAs, resulting in the Pol III promoter in a retroposed 5S rRNA copy presumably remains silent or is expressed at a low level [5]. In different species, the copy numbers of the same origin SINE is different. The copy numbers of MsSE1 and MsSE2 in M. sexta and ShSE1 in S. frugiperda were 7513, 16,157 and 11,117, respectively, whereas only 4521 copies of ObSE1 and 863 copies of ObSE2 were found in O. brumata. The genome sizes of M. sexta and S. frugiperda are around 400 Mb, while O. brumata has larger genome size of 618 Mb. Hence, SINE copy number may not correlate with genome size. Some factors of 3’-tail, such as poly(A) tail or short direct repeats length, sequence conservation and distance to the transcriptional terminator, may affect the transposition efficiency of the SINE families [46, 47]. In this study, the varied 3’-tail of these SINEs in different species may have affected their distribution in the genome. However, their relationship with the number of copies cannot be determined at this time.

Based on the divergence of the copies from the consensus sequence, the relative age distribution of identified SINEs was analyzed. Scattered age profiles were found in most SINEs among all species or within the genus, suggesting that the activity and accumulation of these SINEs are dynamic processes that can vary considerably between host lineages and SINE lineages. Especially, the highly identity and concentrated PxSE1 showed that it most likely is a relatively young retroposon in the genome of P. xylostella and was generated by recent explosive amplification. The scattered distribution of PxSE5 copies also suggests that it is older than other SINEs.

SINEs contribute to DBM genome evolution

The ability of TEs to replicate and move in the genome affects the genomic structure, gene expression, and the divergence and evolution of host species [48–51]. The genome size of DBM is 343.575 Mb, of which the intronic region occupies of 35.23% (121.039 Mb) [52]. The integration pattern analysis revealed that the numbers of PxSINEs inserted into introns accounted for 44–51%, only 2–8 copies were inserted into exons, indicating that PxSINEs prefer to insert or accumulate in introns of genic regions. However, the proportions of different SINEs located within introns of Solanaceae range from 15 to 54% [53] and 96% of SINEs inside genes were located inside introns in Zoxis japonica and maize [54], suggesting that the distribution characteristics of SINEs varied in different species. Introns have long been an exemplar of regulated splicing, which affects and enhances almost every step of mRNA metabolism by the act of their removal [55]. In mice, a recent insertion of MT-C retrotransposon into DICER intron truncated its first 6 exons, providing an alternative promoter and a novel first exon. This change resulted in acquisition of oocyte-specific expression and is essential for fertility [56]. We speculate that the insertions of PxSINES into introns may provide signals for alternative splicing and polyadenylation, which may be a reflection of the host response to an ever-changing environment.

Importantly, we also noticed that only 25 copies of SINEs inserted into the genic exons of DBM, of which 13 copies were found in 3’ UTR. In eukaryotic cells, some proteins (such as PUF protein) can bind to regulatory elements in the 3’ UTR of mRNAs and control mRNA stability, translation and localization [57]. The genes with the insertion of SINEs into exons are mainly annotated in terms of metabolism, cell division, signal transduction and transportation, and it remains to be elucidated whether some of the SINE insertions have an influence on gene expression.

HTT of SINEs

Increasing evidence showed that HTT is a common phenomenon. So far, no less than 5689 HTT events have been recorded [58]. However, only a few HTT events of
SINE have been detected, including the Smal-cor SINE between coregonid and common ancestor of salmonid (Hamada et al. 1997), Sauria SINE between reptiles and mammals [59], HaSE2 SINE between Aphis gossypii and Lepidoptera insects [33]. The long-term vertical inheritance property inherent in SINE and its dependence on active partner LINEs to move in new hosts may be the reason why HTT events rarely occur [47, 60], as was confirmed by the partial congruence between the phylogenetic trees of PxSE3 families and host species in this study. Interestingly, SINVPS1E1, a SINE copy inserted into the baculovirus, shared more than 90% identity to the consensus sequence of SjSE1, SlittSE1 and SlituSE1 (Additional file 4: Figure S4B). In addition, the absence of target site duplication as well as upstream host sequence in SINVII, suggested that non-homologous end-joining of double-strand breaks might be the mechanism of HTT, SINV can successfully infect S. litura and S. exigua [61]. S. exigua multicapsid nucleopolyhedrovirus (SeMNPV) DNA can also replicate in five non-permissive cell lines including SF21AEII, CLS-79, Spli-221, hi-5 and BmN4 [62], indicating a wider host range of NPV. Thus, our finding suggests the occurrence of HTT of PcSE1 between baculovirus and Lepidoptera insects. This is not surprising, because population genomics supported baculoviruses as vectors of horizontal transfer of insect transposons [63]. Similarly, the HTT of Helictron transposon Hel-2 and Tc1-like transposon TCp3.2 between insects and associated baculoviruses has been detected [64, 65]. Recent studies have revealed that the occurrence of HTT generally exhibits species ecological relationships, such as host-parasite [66, 67] and predator-prey [68, 69]. Additionally, proviruses have been reported as vectors for HTT of Sauria SINE from reptiles to mammals [59]. Hence, it is necessary to further explore the HTT events of PcsE1-like elements mediated by baculoviruses.

Conclusions

In this study, we identified three tRNA-derived SINEs and two 5S RNA-derived SINEs in the genome of P. xylostella, among which PcsE1 is a relatively young retrotransposon and was generated by recent explosive amplification. Homology searches revealed scattered distribution of these elements in other Lepidopteran insects with variable copy numbers. The preference of PxsSines to insert or accumulate in introns of genic regions indicated that P. xylostella SINE families contribute to structural variation in introns. The identification of PcsE1-like elements in the baculovirus and related lepidopteran host insects provides evidence of horizontal transfer facilitated by host-parasite interactions. These data may have implications for understanding the evolution and HT mechanisms of SINES.

Methods

Data resources

The 235 publicly available insect databases of WGS assemblies including 33 Lepidoptera insects, EST, nucleotide (Nr/Nt), and TSA from National Center for Biotechnology Information (NCBI) (last accessed November 30, 2018) were used in this study (Additional file 12: Table S2). P. xylostella WGS was downloaded from NCBI [52]. As corresponding gene annotation file, the GFF files GCF_000330985.1 were used.

Database search strategy

To identify SINE candidates, database searches were performed and composed of four steps. Firstly, the known SINE sequences, including tRNA-derived HaSE1 from Helicoverpa armigera [33] and BmSE from B. mori [28], 5S rRNA-derived HaSE3 from H. armigera [33], were used as queries for local blastn in the DBM genome. The sequences of high homology (at least 70% identity over at least 50 bp length to query) as well as 500 bp upstream and downstream flanking regions were extracted using TBtools [70] and analyzed for conserved structural motifs of SINEs such as internal RNA Pol III promoter and TSDs. The consensus sequences of PcsSE1 and PcsSE4 were determined by multiple sequence alignments. Secondly, the consensus sequences of PcsSE1 and PcsSE4 were searched against DBM genome by local BLASTN to identify other potential homologous sequences, and two other tRNA-derived PcsSE2 and PcsSE3 and a 5S rRNA-related PcsSE5 were identified. Thirdly, the 50-bp fragment at 3′-end of SINE families was used as query to search potential partner LINEs, and the LINE, PcsLINE1, related to PcsSE5 was identified. Finally, insect genome databases as well as EST, Nr/Nt and TSA databases from NCBI were searched using consensus sequences of these five SINE families as queries to detect SINES in species other than DBM.

Copy number estimation

To estimate copy number and average divergence of SINES, respective consensus sequences were used to search against related databases (Additional file 12: Table S2). All contiguous sequences with at least 80% identity at the nucleotide level to the consensus over 100 bp were used to estimate copy number in all species [71, 72]. Given the high sequence identity of 5′-ends in several copies of different SINE families in DBM, all those indistinguishable copies were ruled out. For example, PcsSE1 and PcsSE2 shared high identity of 120 bp sequence at 5′-ends, thus all copies aligned only with part or all of this 120 bp region in the consensus sequence were excluded for copy number analysis. Further, all fragments sharing at least 80% identity over at least 80% of the length of the consensus sequence were
aligned and used for calculation of average divergence to consensus sequence with Kimura-2 parameter model [73]. The identity value of single copy to consensus sequence was rounded to an integer for the relative age distribution analysis [53].

Gene association and genomic show cases
The association of DBM SINEs with annotated genes were investigated using custom Perl script from MapGene2Chrom (http://mg2c.iask.in/mg2c_v1.0) [74]. The integration of SINEs into genomic regions including introns, coding and untranslated regions as well as the distances of intergenic copies to the closest neighboring gene were determined as described previously [53]. The number of SINEs within each region was counted and the results were graphically represented using MapGene2Chrom.

Sequence analysis and phylogeny
SINEs’ tRNA-like structure was checked with tRNAscan-SE [75], using mixed model and cove score cut off value = 0.01 as default. Multiple SINE copies were aligned by MUSCLE [76], and the alignments were visualized with GENEDOC (www.psc.edu/biomed/genedoc). The phylogeny of full consensus sequences of SINE families was built by MEGA 7.0 using Maximum Likelihood with K2 + G model [77]. The reliability of the trees was tested using 1000 bootstrap replications [71].

Abbreviations
SINEs: Short interspersed nuclear elements; LINEs: Long interspersed nuclear elements; TE: Transposable elements; RT: Reverse transcriptase; DBM: Diamondback moth; TSDs: Target site duplications; UTR: Untranslated regions; WGS: Whole genome shotgun; EST:Expressed sequence tags; CDS: Coding regions; HTT: Horizontal transposon transfer; TSA: Transcriptome

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07543-z.

Acknowledgements
Not Applicable.

Authors’ contributions
JW and JX conceived and designed the research. GH, NZ, HJ, XM, KQ and YZ analyzed the data. GH and JW wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No. 31701792), Jiangsu Agricultural Science and Technology Innovation Fund (Grant No. ZX (17)2002).

Availability of data and materials
The original sequences used to construct the consensus sequences of all SINEs are uploaded to the NCBI database (http://www.ncbi.nlm.nih.gov/ bioproject/6658855). The complete genome of Plutella xylostella is available at the NCBI RefSeq assembly database under the accession number GCF_000330985.1 and the other lepidopteran genome databases including the accession numbers are available in Additional file 12: Table S2.

Declarations
Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.
45. Kajiwara M, Okada N. LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell. 2002;111(3):433–44. https://doi.org/10.1016/S0092-8674(01)00413-3.

46. Roy-Engel AM, Salem A-H, Oneyranar OJ, Deininger L, Hedges DJ, Klitroy GE, et al. Active Alu element “A-tails”: size does matter. Genome Res. 2002;12(9):1333–44. https://doi.org/10.1101/gr.389802.

47. Conneau MS, Roy-Engel AM, Hedges DJ, Deininger PL. Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? Genome Res. 2009;19(4):545–55. https://doi.org/10.1101/gr.089789.108.

48. Liu D, Yang J, Tang W, Zhang X, Royster CM, Zhang M. SINE retrotransposition drives ecotypic disparity in natural populations of Collia nanus. Mob DNA. 2020;11(1):4. https://doi.org/10.1186/s13100-019-01916-8.

49. Ray DA, Grimshaw RJ, Halsey MK, Korstian JM, Osmanski AB, Sullivan KAM, et al. The Alu sequence provides insights into rapid te-based genome diversification and multiple SINE births and deaths. Genome Biol Evol. 2019;11(8):2162–77. https://doi.org/10.1093/gbe/evz125.

50. Trizzino M, Kapusta A, Brown CD. Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics. 2018;19(1):468. https://doi.org/10.1186/s12864-018-4850-3.

51. Huang X, Li CF, Wu J, Wei IH, Zou Y, Han MJ, et al. Enhancer activity of Helitron in sericin-1 gene promoter from Bombyx mori. Insect Sci. 2016;23(3):396–405. https://doi.org/10.1111/1439-0411.12347.

52. You M, Yue Z, He W, Yang X, Yang G, Xie M, et al. A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet. 2013;45(2):220. https://doi.org/10.1038/ng.2013.206.

53. Cote B, Angra N, Hance AL, K!: Pei B, Tusche J, Schmitt T. Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. Plant J. 2016;85(3):268–85. https://doi.org/10.1111/tpj.13170.

54. Mao H, Wang H. Distribution, diversity, and long-term retention of grass short interspersed nuclear elements (SINEs). Genome Biol Evol. 2017;9(8):2046–56. https://doi.org/10.1093/gbe/evx145.

55. Liu H, Nott A, Moore MJ. How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci. 2003;28(4):215–20. https://doi.org/10.1016/S0968-0004(03)00052-5.

56. Flemr M, Malik R, Franke V, Nejezpiskaj J, Sedlacek R, Vlahovic K, et al. A retrotranspon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. Cell. 2013;155(4):807–16. https://doi.org/10.1016/j.cell.2013.10.001.

57. Wickens M, Banstein DS, Kimble J, Parker R, A. PUF family portrait: 3′ UTR regulation as a way of life. Trends Genet. 2002;18(3):150–7. https://doi.org/10.1016/S0168-9525(02)01026-6.

58. Dotto GP, Carvalho EL, da Silva AF, Dezordi FZ, Pinto PM, Campos TLd, Rezende AM, Wallau GdL. HTT-DB: new features and updates. Database. 2018;2018:2018. https://doi.org/10.1093/database/bax102.

59. Hada K, Misukure Q, Okada N. Proviruses as possible vectors for horizontal transfer of retroelements from reptiles to mammals. Proc Natl Acad Sci U S A. 2007;104(29):12046–51. https://doi.org/10.1073/pnas.0700531104.

60. Luchetti A, Mantovani B. Rare horizontal transmission does not hide long-term inheritance of SINE highly conserved domains in the metazoan genome. Integr Comp Biol. 2020;11(1):4. https://doi.org/10.1186/s13100-019-0165-4.

61. Takatsuka J, Okuno S, Ishii T, Nakai M, Kunimi Y. Host range of two multiple nucleopolyhedroviruses isolated from Spodoptera exigua among insect Lepidoptera. Acta Virol. 1998;42(5):293–9. https://doi.org/10.1016/S0065-1318(05)1318-7.

62. Yanase T, Yasunaga C, Kawarabata T. Replication of Spodoptera exigua nucleopolyhedrovirus in permissive and non-permissive lepidopteran cell lines. Acta Virol. 1998;42(5):293–9.

63. Gilbert C, Chateigner A, Emmerwein L, Barbe V, Bézier A, Herniou EA, et al. Population genomics supports baculoviruses as vectors of horizontal transfer of insect viruses. Nat Commun. 2014;5(1):1–9.

64. Jehle JA, Nickel A, Vlak JM, Backhaus H. Horizontal escape of the novel Tc1–Mariner Helitron from transgenic lettuce. Nat Mol Biol. 1998;46(2):215–24. https://doi.org/10.1016/S0022-2836(98)83594-3.

65. Coates BS. Horizontal transfer of a non-autonomous Helitron among insect and viral genomes. BMC Genomics. 2015;16(10):137. https://doi.org/10.1186/s12864-015-1318-6.

66. Guo X, Gao J, Li F, Wang J. Evidence of horizontal transfer of non-autonomous Lep1 Helitrons facilitated by host-parasite interactions. Sci Rep. 2014;4:5119.

67. Han G, Zhang N, Xu J, Jiang H, Li C, Zhang Z, et al. Characterization of a novel Helitron family in insect genomes: insights into classification, evolution and horizontal transfer. Mob DNA. 2019;10(1):25. https://doi.org/10.1186/s13100-019-0164-5.