三种方法对草地贪夜蛾基因组转座子的注释*

张春辉[1**] 王 磊[1] 刘 运[1] 彭长军[1, 2] 岳碧松[1] 李 静[1***]

（1. 四川大学生命科学学院，生物资源与生态环境教育部重点实验室，成都 610064；2. 中国科学院成都生物研究所，成都 610041）

摘 要 【**目的**】转座子（Transposable element，TE）是昆虫基因组的重要组成，不同昆虫类群的 TE 组成、基因组占比及转座活性等基本特征存在巨大差异。本研究旨在探究不同方法对于草地贪夜蛾 *Spodoptera frugiperda* TE 的注释效果，并在基因组水平阐明草地贪夜蛾 TE 的基本特征。【**方法**】采用 3 种方法对草地贪夜蛾基因组 TE 进行预测，包括基于数据库 Repbase、ArTEdb 进行同源预测，基于重复序列的特性和结构进行从头预测。【**结果**】ArTEdb 方法和从头预测方法鉴定的 TE 分别占基因组 21.48%和 27.26%，其中 LINE 元件无论是拷贝数还是分布密度都最高；其次是 DNA 元件。2 种方法预测的 TE 分歧率分布峰值约 10%，而分歧率<10%的 TE 主要是 DNA 转座子和 LINE。比较 3 种方法的预测结果，Repbase 方法灵敏度低，预测的 TE 远少于其他 2 种方法。ArTEdb 方法能注释出更多的 TE，但该方法对于已知超家族鉴定效果不佳。而从头预测注释出的 TE 数量多，且能划分到不同超家族，甚至能鉴定不包含在 Repbase 鳞翅目库的 TE 超家族。【**结论**】草地贪夜蛾基因组最主要的 TE 类型是 LINE 和 DNA 元件，基因组存在大量年轻的转座子，草地贪夜蛾在 TE 家族的组成上与其它鳞翅目物种存在差异。从头预测的方法对草地贪夜蛾基因组 TE 注释效果较其它 2 种方法更好。这一研究结果为深入研究转座子的功能及其对草地贪夜蛾基因组多样性奠定了基础。

关键词 转座子；Repbase；ArTEdb；从头预测；草地贪夜蛾析

Comparison of three annotation methods for characterizing transposable elements in the *Spodoptera frugiperda* genome

ZHANG Chun-Hui[1**] WANG Lei[1] LIU Yun[1] PENG Chang-Jun[1, 2] YUE Bi-Song[1] LI Jing[1***]

(1. Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610064, China; 2. Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, China)

Abstract **[Objectives]** To find the best method to annotate and characterize Transposable elements (TEs), an important component of insect genomes that vary widely in content, families and transposition activity, among insect taxa, in the *Spodoptera frugiperda* genome. **[Methods]** Three annotation methods were used to identify *S. frugiperda* TEs; Repbase- and ArTEdb- based, homologous prediction and RepeatModeler based *de novo* annotation. **[Results]** The ArTEdb and *de novo* methods predicted TE percentages of 21.48% and 27.26%, respectively. LINEs were the most dominant TEs, both in terms of copy number and density. TEs divergence rates peaked at around 10%, indicating that the majority of TEs have appeared in the *S. frugiperda* genome recently. TEs with divergence rates <10% were mainly LINEs and DNA transposons. In addition, several superfamilies were abundant in *S. frugiperda* that are rare or lacking in other Lepidopteran species. Of the three annotation methods, Repbase had the lowest sensitivity and identified the least number of TEs. ArTEdb identified abundant TEs, and had good sensitivity to DNA elements but failed to further classify TEs into superfamilies. The *de novo* method not only identified the most TEs but also successfully classified these into different superfamilies. It also identified new superfamilies currently not included in the Lepidopteran repbase database. **[Conclusion]** We successfully characterized TE

*资助项目 Supported projects：四川省科技厅重点研发项目（2019YFN0180）

**第一作者 First author，E-mail：976655025@qq.com

***通讯作者 Corresponding author，E-mail：ljtjf@126.com

收稿日期 Received：2020-08-12；接受日期 Accepted：2021-02-06

content and composition in *S. frugiperda* and found that the *de novo* method was superior to the Repbase and ArTEdb methods in terms of both identifying and categorizing TEs. These findings improve our understanding of the TEs of *S. frugiperda* and should benefit further studies on the functional significance of TEs and their contribution to genomic diversity.

Key words transposable element; Repbase; ArTEdb; de novo; *Spodoptera frugiperda*

草地贪夜蛾 *Spodoptera frugiperda* 是一种鳞翅目、夜蛾科、灰翅夜蛾属的农业害虫，源产于美洲的热带和亚热带地区，自 2018 年 12 月首次在云南发现后，目前已迅速扩散到我国广泛的地区（Sun *et al.*，2019）。草地贪夜蛾具有强大的迁飞能力，成虫一晚最多可以迁飞 100 km；其食谱极广，已知的寄主植物有 353 种，对玉米的破坏性最强（Montezano *et al.*，2018）。近年来研究者已对草地贪夜蛾的迁飞行为、交配行为、杀虫剂的敏感性和转基因植物的抗性等方面进行了大量研究。在基因组方面，Kakumani 等（2014）首次报道了一个从头组装的草地贪夜蛾基因组，Gouin 等（2017）报道了玉米型和水稻型两个品系的基因组；Liu 等（2019）公布了组装到染色体水平的草地贪夜蛾基因组。这些研究揭示了草地贪夜蛾基因组的基本特征，然而关于其基因组中转座子的组成和特征目前的了解仍非常匮乏。

转座子（Transposable element，TE）是真核生物基因组的重要组成，它们对基因组的结构、基因功能、基因调控等方面有着重要影响（Chenais *et al.*，2012）。根据转座机制不同，TE 可以分为 RNA 转座子和 DNA 转座子两大类。RNA 转座子需要经 RNA 介导，并通过逆转录酶得到 DNA 拷贝，再整合到基因组的新位点。根据是否存在长末端重复（Long terminal repeated，LTR），RNA 转座子可划分为 LTR 和 Non-LTR。Non-LTR 又划分为长散在元件（Long interspersed nuclear elements，LINE）和短散在元件（Short interspersed element，SINE），短散在元件缺少转座相关的酶，必须依靠 LINE 的转座酶，是一种非自主性的转座元件（Zhang and Rong，2012）。DNA 转座子不以 RNA 为媒介，而直接通过转座酶在基因组内自我繁殖（Feschotte and Pritham，2007）。根据编码蛋白序列的排列顺序，以及序列的突变位点，RNA 转座子和 DNA 转座子还可进一步划分为超家族、家族和亚家族等。

研究昆虫转座子目前仍存在较大的挑战。首先，TE 在昆虫基因组中的占比差异巨大。例如同为双翅目的埃及伊蚊 *Aedes aegypti* TE 占基因组约 55%，而南极蠓 *Belgica antarctica* TE 还不到 1%；果蝇属的拟果蝇 *Drosophila simulans* TE 占基因组 10%，嗜凤梨果蝇 *Drosophila ananassae* TE 达到 40%（Petersen *et al.*，2019）。其次，由于进化历史长，不同昆虫类群分化很大，其转座子也存在许多谱系特异的类型。Novosib 超家族存在于许多膜翅目、双翅目物种及家蚕 *Bombyx mori*，但在其他鳞翅目物种中均未发现，而 Zisupton 超家族则仅在鞘翅目动物中存在（Petersen *et al.*，2019）。同时，在 TE 组成上不同昆虫也存在较大差异，红带袖蝶 *Heliconius melpomene* 的 TE 以 DNA 转座子为主，而家蚕的 TE 则主要由 LINE 组成（Kawamoto *et al.*，2019）。此外，TE 的转座活性方面，昆虫基因组中存在许多仍然具有转座活性的年轻转座子，它们在基因组中不断扩张，是昆虫基因组多样性的重要来源。家蚕基因组存在大量分歧率<10%的 TEs，是新近整合到其基因组的元件，从而造成家蚕与其他 3 种鳞翅目物种的显著区别（Wu and Lu，2019）。

由于昆虫 TE 在组成和类型上的复杂性和多样性，对其鉴定和注释一直非常困难。Petersen 调查了 62 种节肢动物的转座子，大部分非模式物种基因组中 TE 仍然属于未知类型（Talla *et al.*，2017；Petersen *et al.*，2019）。尽管目前草地贪夜蛾基因组已组装好，但关于其中 TE 的注释、鉴定、基本特征却远未完成。重复序列注释方法大致分为 2 种，根据同源性进行预测和从头预测（*de novo*）（Thung *et al.*，2014）。前一种方法是利用已知的重复序列数据库对物种进行重复序列注释。其中重复序列数据库共分为 3 类：以 TE 为中心、以基因组为中心和以多态性

为中心（Goerner-Potvin and Bourque，2018）。从头预测方法不依赖于已有的转座子数据库，而利用重复序列或转座子自身的序列或结构特征，构建从头预测算法对序列进行识别，可分为基于拷贝数注释和基于结构注释。前者是根据 TE 的重复特性进行查找（Arkhipova，2017）；后者则利用 TE 的结构进行查找，但很多的 TE 结构复杂，没有一致性的结构，因此适合 LTR 和 MITE 等结构特征明显的元件的查找（Lerat *et al.*，2019）。从头预测更适合于 TE 报道较少的非模式动物的 TE 研究。本研究中拟选择以 TE 为中心的 Repbase 和以基因组为中心的 ArTEdb 2 种同源预测的方法，以及采用 RepeatModeler 构建一致性序列库的从头预测方法，共比较 3 种方法对草地贪夜蛾基因组转座子的鉴定和注释效果，旨在探索草地贪夜蛾基因组 TE 的鉴定方法，揭示草地贪夜蛾转座子的基因组特征，从而为深入研究转座子的功能及其对基因组多样性的影响奠定基础。

1　材料与方法

1.1　基因组数据

从 CNGBdb（https://db.cngb.org/search/project/CNP0000513/）下载草地贪夜蛾全基因组序列，其组装水平为染色体，31 条染色体总长 461.19 Mb，21 809 条 unplaced_scaffold 总长 82.46 Mb。

1.2　实验方法

1.2.1　基于 Repbase 进行同源预测　Repbase（https://www.girinst.org/downloads/）是以 TE 为中心的转座子数据库，是真核基因组通用转座子数据库。下载 RepeatMasker（4.0.7）（http://repeatmasker.org/RMDownload.html）及数据库 Repbase 进行本地安装，选择鳞翅目作为其查询数据库类别，参数设置为“-e crossmatch -gff -Xsmall -s -low -pa 15”，对草地贪夜蛾基因组进行重复序列注释。

1.2.2　基于 ArTEdb 进行同源预测　ArTEdb（http://db.cbi.pku.edu.cn/arte/index.html）是北京大学生物信息中心建立的以基因组为中心的节肢动物转座子数据库，包含 12 种节肢动物重复序列库。下载 ArTEdb 的草地贪夜蛾的重复序列库，使用 Python 脚本修改重复序列格式，以便于 RepeatMasker 进行分类。以修改后的重复序列库为自定义库，对草地贪夜蛾基因组进行重复序列注释。

1.2.3　从头预测方法进行注释　下载 RepeatModeler（http://repeatmasker.org/RepeatModeler/）及相关的软件 RECON，Tandem Repeats Finder，RepeatScout 进行配置，RepeatModeler 自动调用相关软件构建一致性序列库，并进行分类。使用 TRF 对一致性序列进行注释，串联重复占有率大于 25%的一致性序列去除。将过滤后的一致性序列与 nr 数据库进行比对，过滤掉一致性大于 30%，占有率大于 50%的序列。将未知类型和已知类型分为两个文件，使用本地安装的 blast 软件将未知类型序列比对到转座子蛋白质数据库 Tpases020812（http://www.hrt.msu.edu/uploads/535/78637/Tpases020812.gz），将比对上转座子蛋白质的未知类型序列划入相对应的的超家族。使用 LTR_finder（Xu and Wang，2007）和 LTR_harvest（Ellinghaus *et al.*，2008）预测 LTR 转座子，并将两个程序的结果通过 LTR_retriever 进行过滤（Ou and Jiang，2018）。将不同类型的转座子分别使用 usearch 进行聚类去冗余（Edgar，2010），最后将已知类型和未知类型的转座子合并。将合并后的一致性序列库作为自定义库，使用 RepeatMasker 对基因组进行注释。

1.3　数据分析

使用自行编写的 Python 脚本对 RepeatMasker 输出结果进行统计，使用 Excel 软件绘图。countTE.py 脚本对 RepeatMasker 输出的.gff 文件进行分析，统计 3 种方法注释出的 TE 拷贝数和基因组占比，将统计结果保存在 Excel 文件输出。

TE 序列与其一致性序列之间突变的差异，即为转座子的分歧率。差异越大，分歧率就越高，说明该 TE 更古老，插入基因组的时间也更久远（Churakov *et al.*，2010）。编写 analysizeTEKmer.py 脚本，对 RepeatMasker 输出的.cat.gz 文件进行分析，将分析结果保存在 Excel 文件输出，使用

Excel 软件生成分歧率图。

analysizeOutFile.py 脚本对 RepeatMasker 输出的.out 文件进行分析，统计每条染色体和每条 scaffold 上转座子拷贝数。根据染色体和 scaffold 的长度，计算出染色体和 scaffold 上 TE 密度。

使用自行编写的 countCopyNum.py 脚本，统计不同方法中转座子各家族的拷贝数，分别输出到 Excel 文件。使用自行编写的 overlap.py 脚本，以上一步输出的 Excel 文件为输入文件，输出整合结果。

2 结果与分析

2.1 构建草地贪夜蛾重复序列的一致性序列

采用从头预测方法研究重复序列，需要构建草地贪夜蛾基因组重复序列的一致性序列。本研究结合基于拷贝数和基于结构的方法，共构建出 1 146 条一致性序列（表 1）。其中 419 条为已知

表 1 RepeatModeler 构建出草地贪夜蛾已知类型转座子的一致性序列统计

Table 1 Statistics of known consensus sequences of transposable elements built by RepeatModeler

TE 类型 Classes of TEs	超家族 Superfamilies	超家族数量 Number of elements	总数量 Total number
长末端重复序列 Long terminal repeated	LTR/Unknown	18	183
	LTR/Pao	8	
	LTR/Gypsy	127	
	LTR/DIRS	1	
	LTR/Copia	29	
长散在重复序列 Long interspersed nuclear element	LINE/RTE	41	185
	LINE/R1	24	
	LINE/Proto2	2	
	LINE/Penelope	2	
	LINE/L2	42	
	LINE/I	8	
	LINE/Dong	9	
	LINE/CRE	1	
	LINE/CR1	56	
短散在重复序列 Short interspersed nuclear element	SINE/tRNA	12	14
	SINE/5S	2	
DNA 转座子 DNA transposon	DNA/Unknown	8	167
	DNA/Zator	4	
	DNA/TcMar	30	
	DNA/Sola-3	1	
	DNA/Sola-1	5	
	DNA/PiggyBac	4	
	DNA/PIF	13	
	DNA/P	2	
	DNA/MULE	1	
	DNA/Maverick	1	
	DNA/hAT	25	
	DNA/Ginger	1	
	DNA/CMC	4	
	DNA/Academ	5	
	RC/Helitron	64	

类型的 TE，主要为 LINE（44.9%）和 DNA 转座子（41.3%），而 LTR 和 SINE 较少。595 条为未知类型重复序列。

2.2　3 种方法注释转座子的拷贝数

3 种方法对草地贪夜蛾基因组转座子的注释结果见表 2。Repbase 注释的转座子远远低于其它 2 种方法，从头预测方法注释的转座子最多。但由于从头预测构建的一致性序列有 58%属于未知类型 TE，因此其中约一半都属于未知类型的 TE。在已知类型的 TE 中，LINE 和 DNA 转座子的拷贝数和基因组占比都远高于 SINE 元件和 LTR 元件，是最主要的转座子类型。3 种方法对不同类型 TE 的注释能力存在较大差异，其中从头预测对 LINEs 的注释较好，注释的拷贝数超过 33 万，占基因组 10.09%，而其它 2 个方法的 LINE 仅占 4%和 6.24%。ArTEdb 对 DNA 转座子的注释最好，其拷贝数超过 32 万，占基因组 8.30%，而其它 2 种方法注释的 DNA 转座子仅 0.32%和 1.64%。

2.3　3 种方法注释转座子的分歧率

3 种方法对各类型 TE 的分歧率统计结果见图 1。Repbase 预测的 TE 其分歧率的分布与其它 2 种方法存在较大的差异，绝大多数 TE 分歧率约 25%，而 ArTEdb 和从头预测的 TE 分歧率峰值均约 10%（图 1：A）。在分歧率<10%的年轻的转座子中，Repbase 方法几乎都来自 DNA 转座子，即 DNA 转座子是草地贪夜蛾基因组较活跃的 TE 类型；ArTEdb 方法中则以 DNA 元件为

表 2　3 种方法对草地贪夜蛾基因组转座子注释结果统计

Table 2　Statistics of transposable elements in *Spodoptera frugiperda* genome annotated by three methods

TE 类型 Classes of TEs	Repbase 方法 Repbase method	ArTEdb 方法 ArTEdb method	从头预测方法 *de novo* method
短散在重复序列 Short interspersed nuclear element	20 164（0.61%）	15 024（0.49%）	39 456（0.89%）
长散在重复序列 Long interspersed nuclear element	104 918（4.00%）	253 070（6.24%）	336 570（10.09%）
长末端重复序列 Long terminal repeated	3 368（0.34%）	68 833（2.13%）	77 343（1.91%）
DNA 转座子 DNA transposon	11 703（0.32%）	327 155（8.30%）	44 566（1.54%）
未知 Unknown	57 343（1.32%）	188 488（4.32%）	527 597（13.01%）
合计 Total	713 587（6.00%）	852 570（21.48%）	1 025 532（27.26%）

括号内的数字表示各类转座子在基因组中的占比。

The number in parenthesis represents the proportion of each class of transposable elements in the genome.

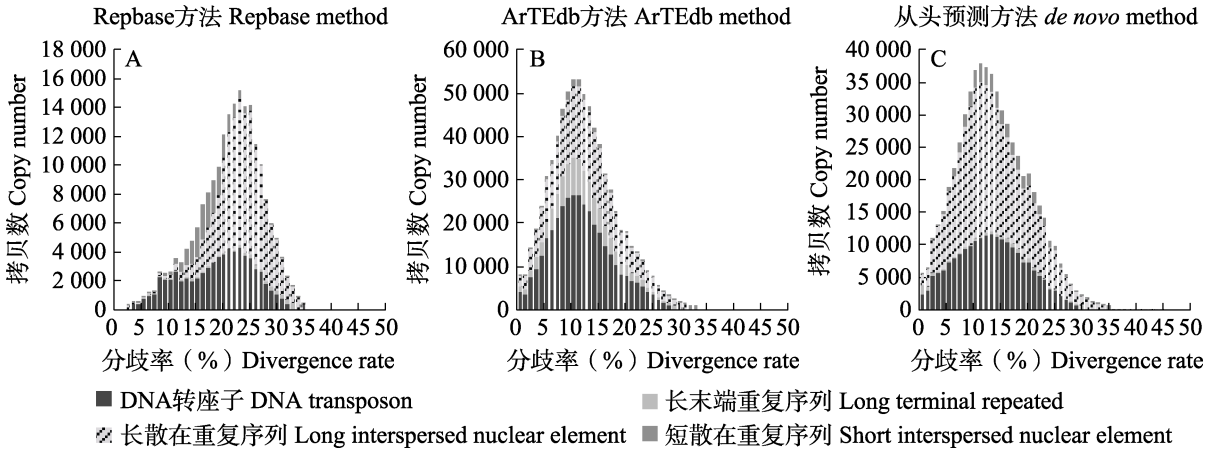

图 1　Repbase 方法（A）、ArTEdb 方法（B）和从头预测方法（C）的草地贪夜蛾基因组已知类型转座子的分歧率

Fig. 1　Divergence rate of known transposable element in *Spodoptera frugiperda* genome annotated by Repbase method (A), ArTEdb method (B) and *de novo* method (C)

主，还包括一定的 LINE 和 LTR（图 1：B）。而从头预测方法，年轻的 TE 主要来自 LINE 元件和 DNA 转座子（图 1：C）。

2.4 3 种方法注释转座子的分布和密度

3 种方法对草地贪夜蛾各染色体和 scaffold 上的 TE 注释结果见表 3，绝大多数 TE 能比对到各染色体，约 1/5 的 TE 分布在 scaffold 上。

草地贪夜蛾 31 条染色体上 TE 的密度分布见图 2。虽然不同方法鉴定的 TE 密度差异很大，但它们在不同染色体上的分布规律大致相似。31 号染色体 TE 密度最高，其次为 28、30 和 13 号染色体；23 号染色体 TE 密度最低（图 2：A）。

各染色体上的 SINE 元件，从头预测方法的密度最高，ArTEdb 方法的密度最小（图 2：B）。LINE 元件，从头预测方法的密度最高，Repbase 方法预测的密度最小（图 2：C）。LTR 元件，则是 ArTEdb 方法的密度远高于其它 2 种方法（图 2：D）。DNA 转座子，ArTEdb 方法的密度最高，Repbase 方法的密度最小（图 2：E）。

对于 scaffold 上的 TE，本文统计了 TE 密度最高的前 20 条 scaffold（图 3），3 种方法鉴定的高 TE 密度的 scaffold 各不相同。Repbase 方法 TE 密度最高为 scaffold9208（5.93 个/kbp），最低为 scaffold21834（3.94 个/kbp）（图 3：A）；ArTEdb 密度最高为 scaffold2640（9.62 个/kbp），最低为 scaffold10044（7.59 个/kbp）（图 3：B）；从头预测密度最高为 scaffold19186（8.39 个/kbp），最低为 scaffold1706（6.43 个/kbp）（图 3：C）。各类型的 TE 在 scaffold 上的分布并不均匀，有的 scaffold 上仅有某一种类型的转座子，如 scaffold20524 上仅有 LINE 元件。这可能是由于大部分 scaffold 长度较短，导致 TE 分布不均的结果。

2.5 3 种方法注释转座子超家族的比较

3 种方法注释草地贪夜蛾 TE 超家族的结果见图 4。对于 SINE 元件（图 4：A），Repbase 和从头预测方法将所有 SINE 元件划分为 2 个超家族：tRNA 和 5S 超家族，而 ArTEdb 方法未鉴定出 5S 超家族；3 种方法均鉴定出 tRNA 超家族是最主要的 SINE 元件。从头预测鉴定的 SINE 元件数量远高于其他两个方法，但未进一步划分到已知的亚家族，而 2 个同源预测方法的结果显示 HaSE2_DP 是数量最丰富的亚家族。

对于 LINE 元件，3 种方法均能鉴定出 9 个超家族（图 4：B）；此外仅 ArTEdb 鉴定了 Jockey 超家族（165 个拷贝），但拷贝数较少。从头预测方法鉴定的超家族拷贝数远远高于其它 2 种方法，尤其是 CR1 和 R1 超家族是最丰富的 LINE 元件，其拷贝数都超过 10 万份，RTE 超家族也超过 5 万份，其他超家族的拷贝则很少。ArTEdb 也能鉴定大量的 LINE 元件，但大部分都未划分到已知超家族。ArTEdb 方法和 Repbase 方法也能够鉴定出一定数量的 RTE，CR1 超家族，但鉴定出的 R1 超家族的拷贝数远不如从头预测方法。

LTR 超家族及其拷贝数的鉴定显示（图 4：C），从头预测和 Repbase 方法都能将 LTR 划分为 4 个不同的超家族；虽然 DIRS 超家族未被

表 3 3 种方法注释草地贪夜蛾染色体和 scaffolds 上的重复序列统计

Table 3 Statistics of transposable elements in *Spodoptera frugiperda* genome scaffolds and chromosomes annotated by three methods

序列类型 Sequence type	方法 Method					
	Repbase 同源预测 Repbase method		ArTEdb 同源预测 ArTEdb method		RepeatModeler 从头预测 *de novo* method	
	染色体 Chromosome	骨架序列 Scaffold	染色体 Chromosome	骨架序列 Scaffold	染色体 Chromosome	骨架序列 Scaffold
被注释的序列数量 Number of annotated sequences	31	15 315	31	19 590	31	20 718
拷贝数 Copy number	160 424	44 601	572 181	144 166	533 291	131 786

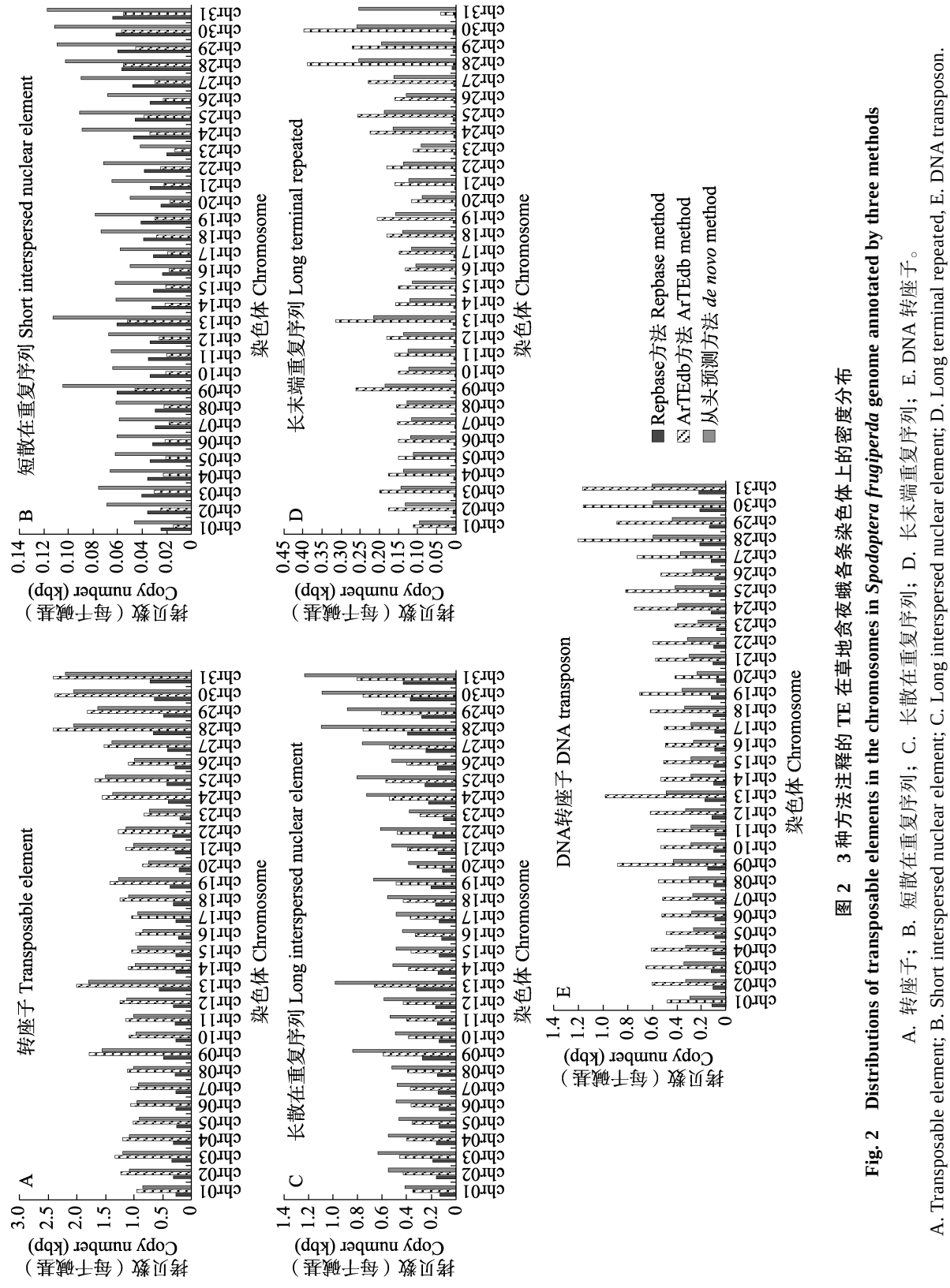

图 2　3 种方法注释的 TE 在草地贪夜蛾各条染色体上的密度分布

Fig. 2　Distributions of transposable elements in the chromosomes in *Spodoptera frugiperda* genome annotated by three methods

A. 转座子；B. 短散在重复序列；C. 长散在重复序列；D. 长末端重复序列；E. DNA 转座子。

A. Transposable element; B. Short interspersed nuclear element; C. Long interspersed nuclear element; D. Long terminal repeated, E. DNA transposon.

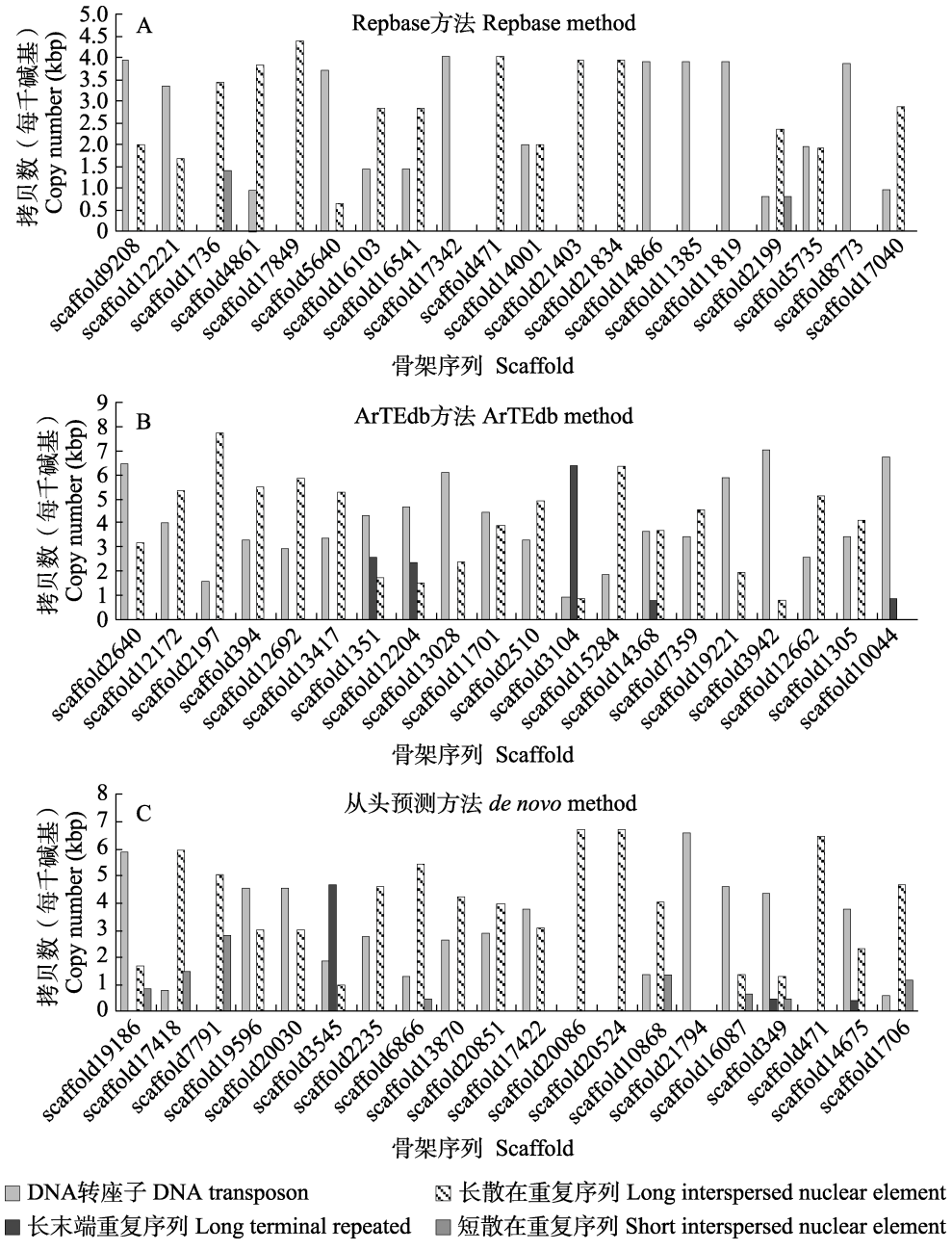

图 3　Repbase 方法（A）、ArTEdb 方法（B）和从头预测方法（C）注释 TE 密度最高的前 20 条草地贪夜蛾 scaffolds 上 TE 的分布

Fig. 3　Distributions of transposable elements in the top 20 TE-densest scaffolds in *Spodoptera frugiperda* genome annotated by Repbase method (A), ArTEdb method (B) and *de novo* method (C)

ArTEdb 方法鉴定出，但使用其余 2 种方法鉴定出的该家族的数量也很低。从头预测对 Gypsy（11 638 个，占总 LTR 的 13%）和 Copia（13 539 个，占总 LTR 的 16%）超家族的鉴定效果较好，ArTEdb 对 Pao 超家族鉴定较好（11 244 个，占总 LTR 的 15.3%）；此 2 种方法鉴定出的 LTR 元件总数量虽然远高于 Repbase 方法，但大部分的 LTR 元件未划分至已知的超家族中。

DNA 超家族和拷贝数的鉴定显示（图 4：D），从头预测方法鉴定出 DNA 超家族最多（13 个），拷贝数也较多。ArTEdb 方法鉴定出的 DNA 转座子的拷贝数最多，但其超家族种类也最少，仅为 6 个。Repbase 方法能够鉴定出较多种类的 DNA 超家族（9 个），但其拷贝数是最少的。虽

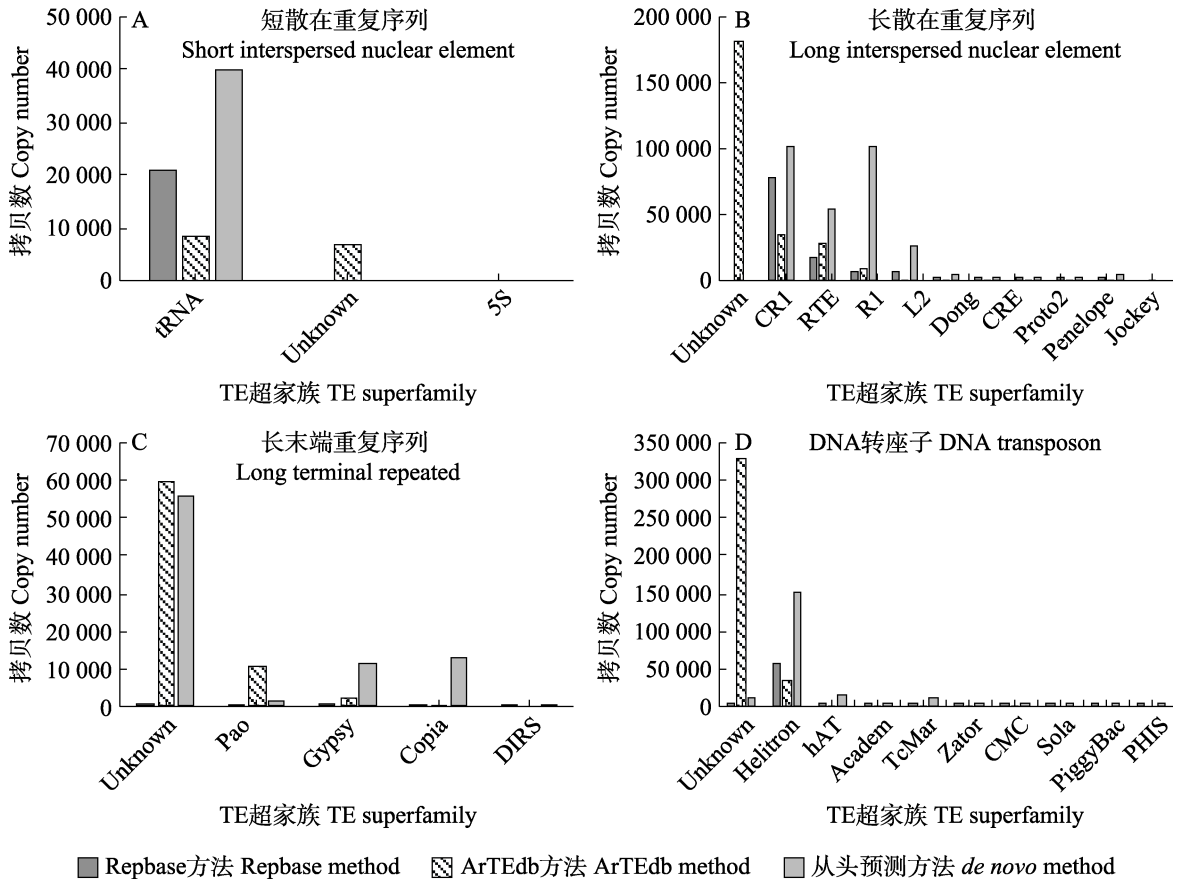

图 4　3 种方法中草地贪夜蛾基因组中各转座子家族注释统计

Fig. 4　Copy number of various types of TE superfamilies in *Spodoptera frugiperda* genome annotated by 3 methods

A. 短散在重复序列；B. 长散在重复序列；C. 长末端重复序列；D. DNA 转座子。

A. Short interspersed nuclear element; B. Long interspersed nuclear element; C. Long terminal repeated; D. DNA transposon.

然 3 种方法鉴定的拷贝数各有差别，但各方法识别出拷贝数最多的 DNA 超家族均为 Helitron 超家族（拷贝数>10 000）。此外从头预测方法鉴定出另外 2 个拷贝数较多的超家族，分别为 hAT 超家族（14 771 个）和 TcMar 超家族（10 695 个）。

3　讨论

关于草地贪夜蛾基因组的转座子鉴定，过去的报道差异较大。Kakumani 等（2014）预测 TE 占其基因组 19.39%，而基因组 18.21%都属于未知类型 TE。Gouin 等（2017）推测 TE 占草地贪夜蛾基因组 29%，其中主要为 SINE 元件（12.5%）。Wu 和 Lu（2019）则认为草地贪夜蛾 TE 占基因组 19.02%，其中主要是 DNA 转座子（9.07%）和 LINE（5.1%）。对同一个物种的 TE 鉴定结果差异如此大，一方面是由于不同研究使用了不同的参考基因组及不同的 TE 预测方法（Kawamoto *et al.*，2019），另一方面也说明昆虫 TE 注释难度较大（Piegu *et al.*，2015；Bourque *et al.*, 2018; Goerner-Potvin and Bourque, 2018）。本研究通过 3 种方法对草地贪夜蛾基因组 TE 进行了更全面地鉴定和注释，发现除 Repbase 方法外，TE 占比约占草地贪夜蛾基因组 21.48%-27.26%（表 2），这一比例较鳞翅目的家蚕 TE（46.07%）低，而较黑脉金斑蝶 *Danaus plexippus*（13.1%）高（Zhan *et al.*，2011）。草地贪夜蛾 TE 中，无论是拷贝数还是分布密度，LINE 元件都是最高的，是最主要的 TE 类型。此外 ArTEdb 方法还发现 DNA 元件的拷贝数超过 32 万份，也是其重要的 TE 组成。本文统计多篇文献报道的鳞翅目物种的 TE 占比（Zhan *et al.*，2011；Talla *et al.*，2017；Petersen *et al.*，

2019），目前已报道的鳞翅目物种各类 TE 占比为 SINE：0.4%-11.85%，LINE：0.4%-17.49%，DNA 转座子：0.2%-17.13%，LTR：0.16%-5.29%。与已报道的鳞翅目物种 TE 占比相比，草地贪夜蛾的 SINE 元件的含量偏低（0.49%-0.89%），其它类型 TE 大致相当。

统计转座子的分歧率分布可以了解不同类型 TE 整合到基因组的时间（Churakov *et al.*，2010）。草地贪夜蛾 TE 的分歧率基本呈正态分布，2 种方法鉴定的 TE 分歧率分布的峰值仅 10%，表明草地贪夜蛾存在大量新近插入基因组的年轻转座子。在分歧率<10%的年轻 TE 中，主要是 DNA 转座子及一定数量的 LINE 元件（图 1）。说明这 2 种元件是较为活跃的转座子，可能对草地贪夜蛾基因组的结构和功能存在重要的影响。因此，深入研究这些活跃的 TE 元件，有可能揭示它们对草地贪夜蛾独特的行为和生理特性的影响。与之相似，家蚕基因组中也存在大量年轻的 TE，绝大多数 TE 的分歧率<5%，这些大量的年轻、新近插入的 TE 主要来自 DNA 转座子、LTR 和 LINE，它们的扩张被认为可能是家蚕拥有较大基因组（530 Mb）的重要原因（Wu and Lu，2019）。

在 TE 的超家族组成上，本研究结果发现草地贪夜蛾也与其它鳞翅目物种不同。基于组装得更完善的基因组和更好的注释方法，研究表明一些基因组在草地贪夜蛾特别丰富而在其它鳞翅目物种中较少，甚至完全缺乏 TE 家族。如 LINE 中的 CR1-Zenon 家族，在草地贪夜蛾中拷贝数超过 15 万，而该家族仅在其它 4 个鳞翅目物种有报道，且拷贝数都<5 000。DNA 元件中的 hAT 超家族在草地贪夜蛾也很丰富（14 771 个），而其它鳞翅目物种极少。SINEs 中的 5S-Deu 家族，也仅在其他 2 个物种 *Lerema accius* 和庆网蛱蝶 *Melitaea cinxia* 中报道（Talla *et al.*，2017）。甚至 LINEs 中的 Dong-R4 家族仅在草地贪夜蛾中发现。这些特定的 TE 家族在草地贪夜蛾基因组中的扩增可能是造成其与其他鳞翅目物种差异的重要原因之一。

Repbase 方法和 ArTEdb 方法都是基于数据库对基因组 TE 进行同源预测的研究方法。Repbase 是经典的重复序列库，使用该数据库预测的 TE 信息完整，大部分的拷贝能够细分到亚家族（Bao *et al.*，2015）。但在昆虫 TE 研究中，由于 TE 在不同类群中保守性弱，近缘物种之间的 TE 差距都可能非常大，因此 Repbase 方法预测草地贪夜蛾 TE 占比和拷贝数远少于其它 2 种方法，且灵敏度低。ArTEdb 是新近发布的一个包含了 12 种节肢动物 TE 的一致性序列的数据库（Wu and Lu，2019），预测的结果较 Repbase 更加灵敏，能注释出更多的草地贪夜蛾 TE，使得 TE 比例从 6%上升到 21.48%，但该方法无法鉴定新的转座子家族，同时，对于已知超家族的鉴定上，ArTEdb 的鉴定效果也不佳，绝大多数 LINE、SINE、DNA 元件和 LTR 都未能划分到特定的超家族。

本研究通过结合拷贝数和结构的预测方法，利用 RepeatModeler 对草地贪夜蛾 TE 进行从头预测，取得了较好的效果。鉴定的重复序列占基因组 28.33%，（其中 TE 占 27.26%），这与新近报道的草地贪夜蛾基因组的估计结果接近（28.24%）（Liu *et al.*，2019）。该方法尤其是对 LINE、SINE 元件的预测效果最好，不仅预测的数量是 3 种方法中最高的，还能进一步将其划分到不同的超家族。虽然该方法对 LTR 和 DNA 转座子预测的数量不及 ArTEdb 方法，但 ArTEdb 并不能将这些 LTR 和 DNA 转座子划分到特定的超家族。此外从头预测方法还鉴定了 3 种未包含在 Repbase 鳞翅目中的 DNA 转座子超家族（拷贝数大于 100），分别为 DNA/P、DNA/MULE 和 DNA/PHIS 超家族，是一种更灵敏和准确鉴定草地贪夜蛾 TE 的方法。需要注意的是，从头注释的方法注释出的 TE 边界不是非常准确，且可能将高拷贝的基因误判为 TE，需要对构建出的一致性序列进行过滤（Bergman and Quesneville，2007；Ou *et al.*，2019）。如需确定一致性序列所属家族或者划分出新的家族或亚家族，可以按照“80-80-80”规则来进行确定家族或亚家族，即序列比对，序列覆盖度>80%，序列一致性>80%（Wicker *et al.*，2007）。无论是同源预测还是从头预测的方法，本研究对草地贪夜蛾 TE 的鉴定和注释结果显示仍然存在大量未知类型的 TE，

排除部分假阳性的结果，其中可能包含草地贪夜蛾特有的 TE 类型。因此结合多种注释方法，深入研究这些未知类型的 TE，将有可能揭示其对草地贪夜蛾基因组结构及其独特行为和生理功能的影响。

参考文献 (References)

Arkhipova IR, 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mobile DNA*, 8(1): 19.

Bao WD, Kojima KK, Kohany O, 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1): 7.

Bergman CM, Quesneville H, 2007. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, 8(6): 382–392.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C, 2018. Ten things you should know about transposable elements. *Genome Biology*, 19(1): 199.

Chenais B, Caruso A, Hiard S, Casse N, 2012. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1): 7–15.

Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, Schmitz J, 2010. A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. *BMC Evol. Biol.*, 10(1): 376.

Edgar RC, 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19): 2460–2461.

Ellinghaus D, Kurtz S, Willhoeft U, 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9(1): 18.

Feschotte C, Pritham EJ, 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.*, 41(1): 331–368.

Goerner-Potvin P, Bourque G, 2018. Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, 19(11): 688–704.

Gouin A, Bretaudeau A, Nam K, Gimenez S, Aury JM, Duvic B, Hilliou F, Durand N, Montagne N, Darboux I, Kuwar S, Chertemps T, Siaussat D, Bretschneider A, Mone Y, Ahn SJ, Hanniger S, Grenet AG, Neunemann D, Maumus F, Luyten I, Labadie K, Xu W, Koutroumpa F, Escoubas JM, Llopis A, Maibeche-Coisne M, Salasc F, Tomar A, Anderson AR, Khan SA, Dumas P, Orsucci M, Guy J, Belser C, Alberti A, Noel B, Couloux A, Mercier J, Nidelet S, Dubois E, Liu NY, Boulogne I, Mirabeau O, Le Goff G, Gordon K, Oakeshott J, Consoli FL, Volkoff AN, Fescemyer HW, Marden JH, Luthe DS, Herrero S, Heckel DG, Wincker P, Kergoat GJ, Amselem J, Quesneville H, Groot AT, Jacquin-Joly E, Negre N, Lemaitre C, Legeai F, d'Alencon E, Fournier P, 2017. Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda*, Noctuidae) with different host-plant ranges. *Sci. Rep.*, 7(1): 11816.

Kakumani PK, Malhotra P, Mukherjee SK, Bhatnagar RK, 2014. A draft genome assembly of the army worm, *Spodoptera frugiperda*. *Genomics*, 104(2): 134–143.

Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, Fujiyama A, Kiuchi T, Yamamoto K, Shimada T, 2019. High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.*, 107: 53–62.

Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour AB, Vieira C, Gonzalez J, 2019. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol. Ecol.*, 28(6): 1506–1522.

Liu H, Lan T, Fang D, Gui F, Wang H, Guo W, Chen X, Chang Y, He S, Lyu L, Sahu SK, Chen L, Li H, Liu P, Fan G, Liu T, Hao R, Lu H, Chen B, Zhu S, Lu Z, Huang F, Dong W, Dong Y, Kang L, Yang H, Sheng J, Zhu Y, Liu X, 2019. Chromosome level draft genomes of the fall armyworm, *Spodoptera frugiper*da (Lepidoptera: Noctuidae), an alien invasive pest in China. *BioRxiv*, doi: org/10.1101/671560.

Montezano DG, Specht A, Sosa-Gomez DR, Roque-Specht VF, Sousa-Silva JC, Paula-Moraes SV, Peterson JA, Hunt TE, 2018. Host plants of *Spodoptera frugiperda* (Lepidoptera: Noctuidae) in the Americas. *African Entomology*, 26(2): 286–300.

Ou SJ, Jiang N, 2018. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, 176(2): 1410–1422.

Ou SJ, Su WJ, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB, 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, 20(1): 275.

Petersen M, Armisen D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B, 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMCc Evolutionary Biology*, 19(1): 11.

Piégu B, Bire S, Arensburger P, Bigot Y, 2015. A survey of transposable element classification systems-a call for a

fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet Evol.*, 86: 90–109.

Sun XX, Hu CX, Jia HR, Wu QL, Shen XJ, Zhao SY, Jiang YY, Wu KM, 2019. Case study on the first immigration of fall armyworm *Spodoptera frugiperda* invading into China. *Journal of Integrative Agriculture*, doi: 10.1016/S2095-3119(19)62839-X.

Talla V, Suh A, Kalsoom F, Dinca V, Vila R, Friberg M, Wiklund C, Backstrom N, 2017. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (Leptidea) butterflies. *Genome Biology and Evolution*, 9(10): 2491–2505.

Thung DT, de Ligt J, Vissers LEM, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY, 2014. Mobster: Accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology*, 15(10): 488.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH, 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12): 973–982.

Wu CC, Lu J, 2019. Diversification of transposable elements in arthropods and its impact on genome dvolution. *Genes*, 10(5): 338.

Xu Z, Wang H, 2007. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35(Suppl. 2): W265–W268.

Zhan S, Merlin C, Boore JL, Reppert SM, 2011. The monarch butterfly genome yields insights into long-distance migration. *Cell*, 147(5): 1171–1185.

Zhang L, Rong YKS, 2012. Retrotransposons at *Drosophila telomeres*: Host domestication of a selfish element for the maintenance of genome integrity. *Biochimica et Biophysica Acta-Gene Regulatory Mechanisms*, 1819(7): 771–775.