*Zoological Systematics*, 43(4): 341–355 (October 2018), DOI: 10.11865/zs.201832

ORIGINAL ARTICLE

# Conserved profiles of digestion by double restriction endonucleases in insect genomes facilitate the design of ddRAD

Bingyan Li[1, 2], Qiao Gao[2], Lijun Cao[2], Ary Anthony Hoffmann[3], Qiong Yang[3], Jiaying Zhu[1 *], Shujun Wei[2 *]

[1]*Key Laboratory of Forest Disaster Warning and Control of Yunnan Province, College of Forestry, Southwest Forestry University, Kunming 650224, China; E-mails: 1055167220@qq.com, jyzhu001@gmail.com*

[2]*Institute of Plant and Environmental Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China; E-mails: 15100172515@163.com, gmatjhpl@163.com, shujun268@163.com*

[3]*School of BioSciences, Bio21 Institute, The University of Melbourne, Victoria 3010, Australia; E-mails: ary@unimelb.edu.au, qiongy@unimelb.edu.au*

[*]*Corresponding authors, E-mails: jyzhu001@gmail.com, shujun268@163.com*

**Abstract** Double-digested Restriction Site Associated DNA Sequencing (ddRAD) through next-generation sequencing (NGS) generates large numbers of loci for characterizing genome-wide variation among multiple samples using next-generation sequencing. Different combinations of restriction endonucleases (REs) may produce varying size distributions of digested fragments, which affect the number of genotyped loci. Understanding digestion profiles across different species will help in selecting REs for digestion in a particular organism. In this study, we use of genome sequences to compare the *in silico* digestion profile of 26 combinations of REs in 131 insect species with two simulation programs. The number of digested fragments in the 300–450 bp range increases linearly with the size of the genome. Different species and insect orders showed similar profiles when digested by different combinations of REs *in silico*, indicating the conservation of digestion by double enzymes in insect genomes. Combinations with *NlaIII* or *Taq*$^{\alpha}$*I* usually produced higher number of fragments in the range 300–450 bp, while combinations with *EcoRI* or *MluCI* produced fewer fragments. The proportion of fragments with the same overhangs at the two ends of digested DNA was higher than those with different overhangs. The two four-base enzyme pairs produced more fragments in the 300–450 bp range than pairs of four-base + six-base enzymes. Experimental digestion of three species from Hymenoptera, Lepidoptera and Thysanoptera showed profiles congruent with *in silico* expectations. Our results shed light on understanding the digestion profiles of insect genomes and provide guidance on selecting REs for ddRAD projects.

**Keywords** Double-digested RADseq, *in silico* simulation, insect genome, optimal double digestion combination.

## 1 Introduction

The development of next-generation sequencing (NGS) techniques has led to methods for genome-wide discovery and genotyping of thousands of genetic markers for ecological and evolutionary studies (Davey *et al.*, 2011). Restriction-site

Received 1 January 2018, accepted 15 March 2018
Executive editor: Fuqiang Chen

associated DNA sequencing (RADseq) techniques (Miller *et al.*, 2007; Baird *et al.*, 2008) represent a method that use NGS and restriction enzymes (Davey *et al.*, 2013) for population-level genomic comparisons at a reasonable cost not only for model species but also for non-model species lacking genomic information (Andrews & Luikart, 2014; Andrews *et al.*, 2016).

Restriction-site-associated DNA sequencing (RADseq) techniques that have been developed to use different types and numbers of restriction enzymes, methods for size selection and strategies for sample multiplexing, including mbRAD (Miller *et al.*, 2007; Baird *et al.*, 2008), ddRAD (Peterson *et al.*, 2012), 2bRAD (Wang *et al.*, 2012), ezRAD (Toonen *et al.*, 2013), nextRAD (Russello *et al.*, 2015) and quaddRAD (Franchini *et al.*, 2017). Among these methods, ddRAD is the most popular (Peterson *et al.*, 2012; Schweyen *et al.*, 2014; Rašić *et al.*, 2015; Hoffberg *et al.*, 2016; Franchini *et al.*, 2017) and involves two restriction endonucleases (REs) and automated size selection (Peterson *et al.*, 2012; Puritz *et al.*, 2014). The suitable size range of fragments for genotyping depends on the read length of the NGS platform. Enzyme choice, size selection, and sequencing effort have effect on per locus sequence coverage which affects confidence in genotype calls. Polymorphisms identified through ddRAD require that recognition sites of REs are conserved across conspecifics, unless there is a mutation in the cut sites of the REs. This conservation of recognition sites for ddRAD need evaluation across different species.

Current studies with the ddRAD method determine useful combinations of REs empirically through pilot digestion experiments (Yang *et al.*, 2016; Johnson *et al.*, 2017) and/or by simulation when a genome sequence is available. The simulation method is easier when narrowing down the range of suitable RE combinations compared to the experimental method (DaCosta & Sorenson, 2014; Rašić *et al.*, 2014; Mora-Márquez *et al.*, 2017). However, without a reference genome, *in silico* digestion is not possible, and an experimental method may need to be used despite the additional time and cost required. An experimental approach sums the total number of fragments without discriminating ones with the same or different overhangs. Some studies have used combination of REs without prior evaluation based on findings from related species (DaCosta & Sorenson, 2015; Pukk *et al.*, 2015; Derkarabetian *et al.*, 2016; Tigano *et al.*, 2017) but it remains challenging to select suitable RE combinations.

The number of sequenced insect genomes is rapidly increasing, covering 14 orders of insects (Yin *et al.*, 2016), and this provides an opportunity to study predicted digestion profiles based on different combinations of REs across a range of species. In this study, *in silico* digestion profiles by double digestion of the genomes of 131 insect species were generated and compared. Additionally, digestion profiles were empirically generated to validate *in silico* patterns for three representative species.

## 2 Methods

### 2.1 Source of insect genomes for *in silico* digestion

We used 131 published insect genomes from InsectBase (http://genome.zju.edu.cn/) (Yin *et al.*, 2016), covering 14 insect orders (Fig. S1 and Table S1). Of which, orders Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera had the most sequenced genomes.

### 2.2 Selection of restriction endonucleases and simulation

The six- or eight-base pair cutting REs failed to produce large numbers of markers because of rare cut sites (Lowry *et al.*, 2017). We therefore selected seven REs recognizing four-base sequences (*AciI*, *BfaI*, *DpnII*, *MluCI*, *MspI*, *NlaIII* and *TaqαI*) and a commonly-used six-base RE (*EcoRI*) as alternative REs to generate 28 combinations of double digestion REs for testing. Because *MluCI* and *EcoRI* generate the same overhangs, and *BfaI* and *DpnII* have the same recognition sites, combinations of these two REs were not included in analyses.

Four programs have been developed for the simulation of digestion, i.e. *DDsilico* (Rašić *et al.*, 2014), *Digital_RADs.py* of BU-RAD-seq (DaCosta & Sorenson, 2014), SimRAD (Lepais & Weir, 2014) and DDRADSEQTOOLS (Mora-Márquez *et al.*, 2017). The latter three show similar performance in the number of digest fragments identified with different overhangs (Mora-Márquez *et al.*, 2017). We therefore chose *Digital_RADs.py* and *DDsilico* for the simulation of double digestion. Considering the read lengths obtained from NGS sequencing platforms, we counted the number of target fragments with different overhangs from 300 bp to 450 bp among all species and across each of the five orders with the most species.

### 2.3 Statistical analysis

The association between the number of fragments and genome size was analyzed by linear regression based on the hypothesis that fragment number should increase linearly with genome size. We present the pattern for 300 to 450 bp fragments produced by *EcoRI*+*MspI*, the same combination used in the original ddRAD protocol (Peterson *et al.*, 2012).

### 2.4 Experimental digestion of three species

We validated predicted digestion profiles in three species from different orders: *Plutella xylostella* (Lepidoptera: Plutellidae), *Frankliniella occidentalis* (Thysanoptera: Thripidae), and *Apis mellifera* (Hymenoptera: Apidae).

We used a DNeasy Blood & Tissue Kit (Qiagen, USA) to extract genomic DNA from thorax of *P. xylostella* and *A. mellifera* and whole body of *F. occidentalis* and a Qubit 3.0 (Life Invitrogen, USA) to quantify DNA concentration. Based on results of the simulations, four enzyme pairs, *NlaIII*+*BfaI*, *NlaIII*+*AciI*, *NlaIII*+*EcoRI* and *NlaIII*+*MluCI* (NEB, USA) were used for digestion of the extracted genomic DNA. To obtained enough of DNA quality (120 ng), one individual of *P. xylostella*, *A. mellifera*, and ten individuals of *F. occidentalis*, was extracted as one piece of DNA and digested by one enzyme pairs. The samples were digested in a 50 μL reaction volume, with 35 μL of DNA (total of 120 ng DNA) and 15 μL of digestion volume including 8 μL of buffer, 1 μL of each enzyme and 5 μL $H_2O$. To ensure different REs reacting efficiently in one digestion system, CutSmart buffer was used for each combination of REs. Finally, the mixtures were incubated at 37°C for 3 h, and then held at 4°C.

The products of double enzyme digestion were visualized on an Agilent 2100 Bioanalyzer (Agilent Technology, USA) using a high sensitivity DNA chip, to confirm the size distribution of digested fragments. Before using the Agilent 2100 Bioanalyzer, the reaction products were cleaned by 1.5× volume of AMPure XP beads (Beckman Coulter, USA) to remove enzymes and others.

## 3 Results

### 3.1 Correlation between the number of digested fragment and genome size

Overall, regression analysis showed a positive association between the number of digested fragments (in the 300–450 bp range) and genome size ($R^2$ = 0.839), indicating that the species with larger genome size tend to generate more fragments (Fig. 1). Consistent results were generated for analyses when species from Coleoptera (y = 43.69x–804.85, $R^2$ = 0.894), Diptera (y = 54.78x–2585.4, $R^2$ = 0.724) and Hemiptera (y = 17.06x+7170.8, $R^2$ = 0.736), but no association for the

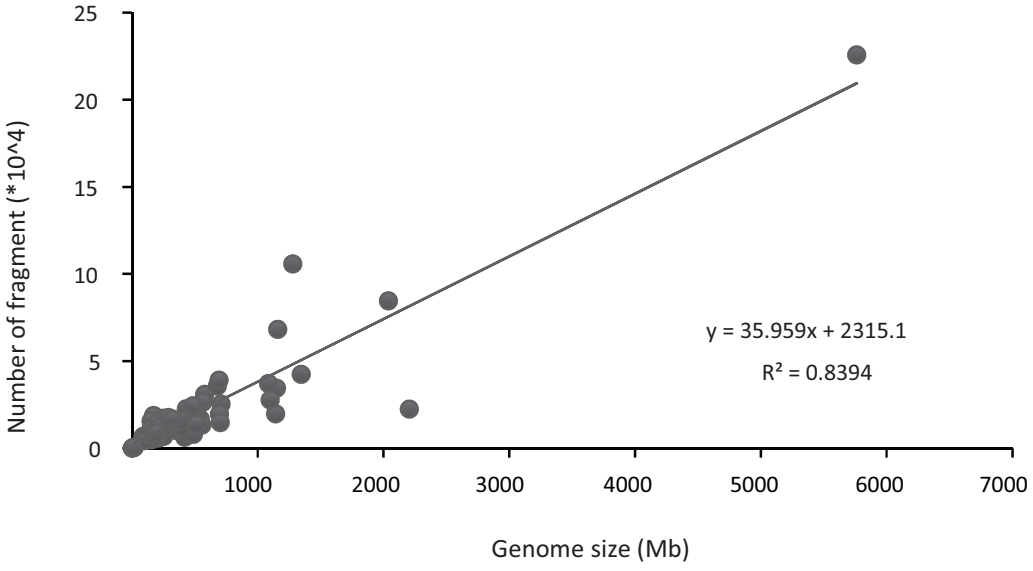


Figure 1. Linear regression of the number of fragments (300–450 bp) digested by *EcoRI* + *MspI in silico* against genome size of 131 insect species.

Hymenoptera ($y=7.43x+12013$, $R^2=0.282$) and Lepidoptera ($y=21.51x+6813.4$, $R^2=0.05$) were included, respectively. To reduce any effect of genome size in subsequent analyses, we used the relative number of fragments by dividing the number of fragments isolated with the genome size of each species.

### 3.2 Ratio between the number of fragments with same and different overhangs

We calculated the number of fragments between 300 and 450 bp having the same overhangs and those having different overhangs, and found that the proportion of fragments with same overhangs was very high, averaging 64.79 % (ranging from 45.92 % to 94.08 %, with ratios having medians up to 4 (Fig. 2). When digested by RE combinations with *EcoRI* or *MspI*, the proportion of fragments with the same overhangs was particularly high, changing the digestion profiles for different combinations of REs (Figs 2, S2, S3), such as *MspI* + *NlaIII*, *EcoRI* + *NlaIII*, *EcoRI* + DpnII and *EcoRI* + *Taq*$^{\alpha}$*I*. In *Dactylopius coccus* and *F. occidentalis*, fragments with the same overhangs were ca. 10 times as common as those with different overhangs.

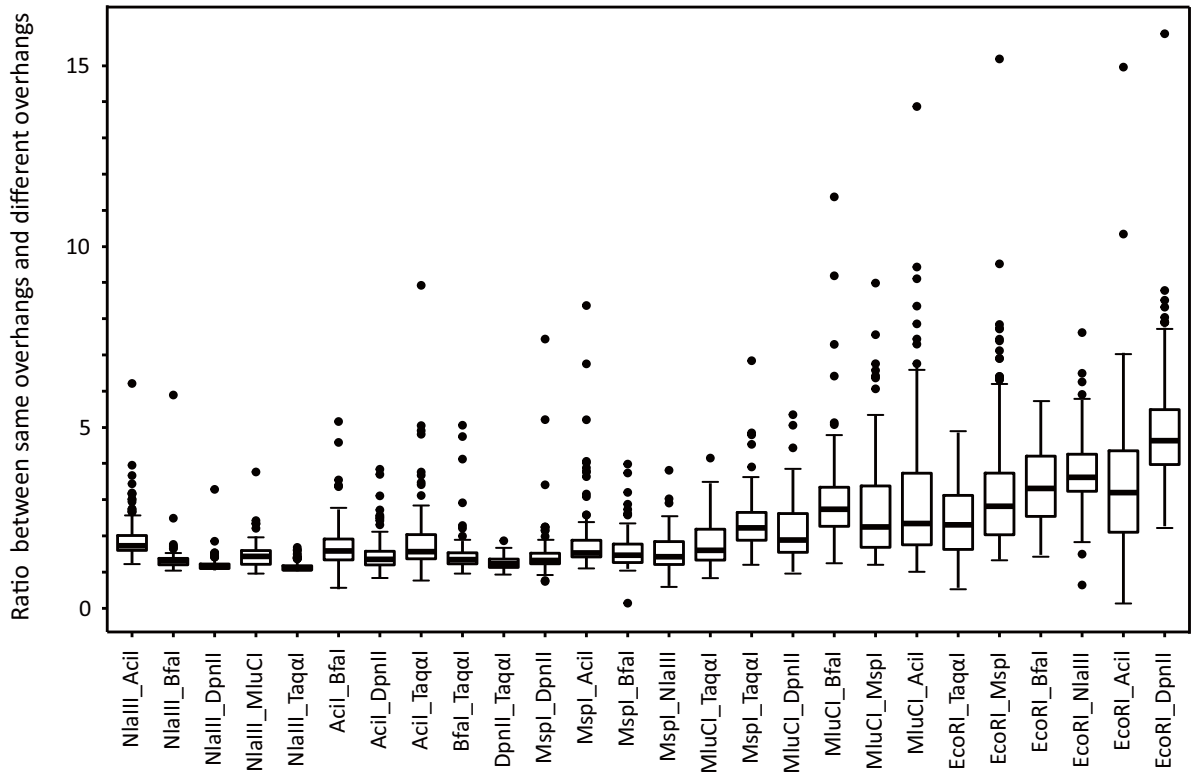


Figure 2. Boxplots for the ratio between the number of fragments with the same and different overhangs for the different combinations of restriction endonuclease pairs. The circles indicate outliers while bars represent medians.

### 3.3 Relative number of fragments with different overhangs digested by combinations of REs

The *DDsilico* analysis generated similar profiles of digestion as the *Digital_RADs.py* and the performance of different combinations of REs was similar across orders and species (Figs 3, S2, S3). The digestion combinations involving *NlaIII* or *Taq*$^{\alpha}$*I* produced a higher number of fragment between 300 bp and 450 bp than with the combination involving *EcoRI* or *MluCI*, with the latter producing half the fragments of the former (Figs 3, S2). When the Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera were analyzed separately, similar results were found except for the Hymenoptera, where combinations involving *DpnII* generated more fragments between 300 bp and 450 bp than those involving *NlaIII* and *Taq*$^{\alpha}$*I* (Fig. 3).

### 3.4 Experimental validation

The average concentration of *P. xylostella*, *A. mellifera* and *F. occidentalis*, was 14.63, 4.01 and 6.32 ng/μl, respectively.

Focusing on fragment sizes from 300 bp to 450 bp, the experimental results were consistent with *in silico* predictions simulated by *DDsilico* for all three species: *P. xylostella* (Fig. 4), *A. mellifera* (Fig. S4) and *F. occidentalis* (Fig. S5). The combinations, *NlaIII*+*BfaI* and *NlaIII*+*AciI* generated more fragments than *NlaIII*+*EcoRI* and *NlaIII*+*MluCI*. However, the distribution of fragments ranging from 150 bp to 1000 bp was different between the *in silico* predictions and experimental results. For *in silico* digestions, most fragments were shorter than 600 bp, while there were few long fragments. In the

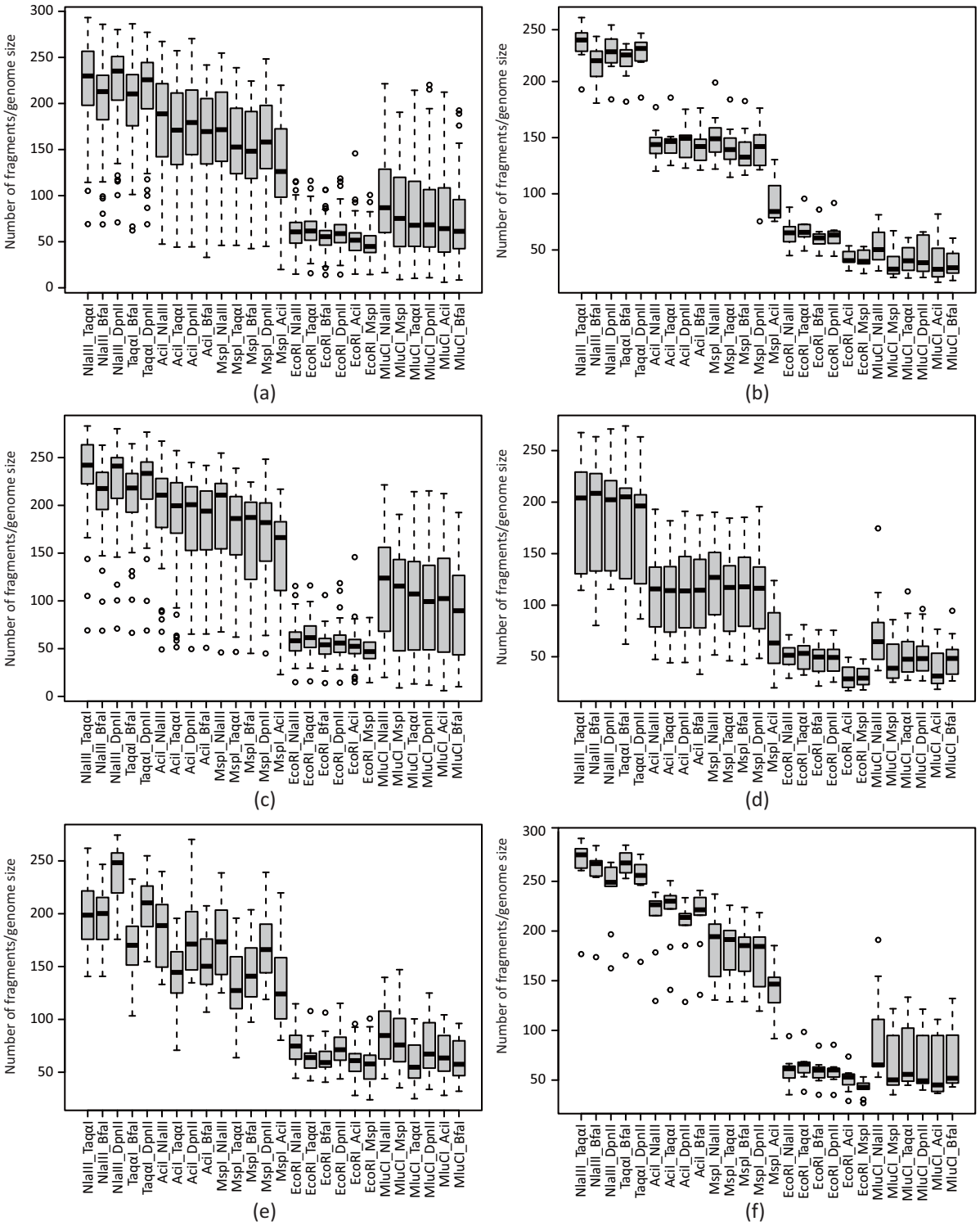


Figure 3. Boxplots for the number of fragments (300–450 bp) with different overhangs digested by 26 combinations of restriction endonucleases simulated by *DDsilico*. (a) all species, (b) Coleoptera, (c) Diptera, (d) Hemiptera, (e) Hymenoptera, and (f) Lepidoptera. The circles indicate outliers. The x-axes show different combinations of restriction endonuclease pairs.

experimental digestions of *NlaIII*+*EcoRI*, there were many more long fragments around 1000 bp.

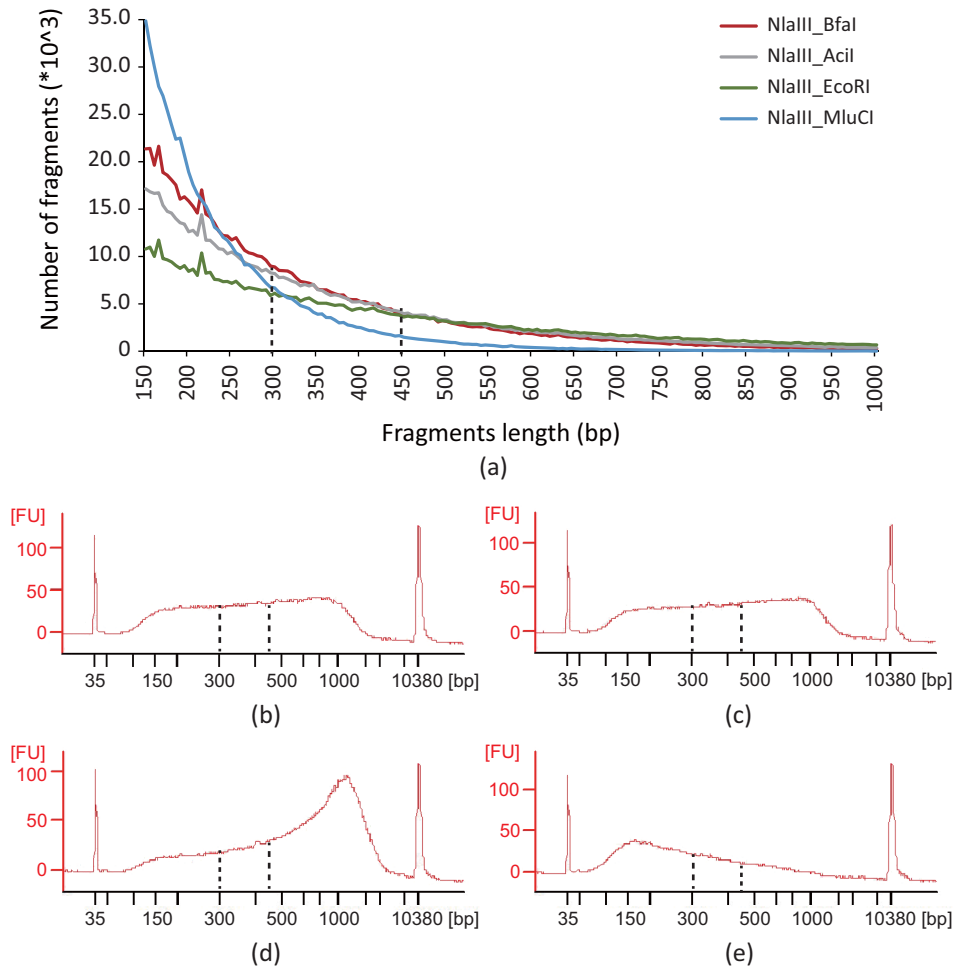


Figure 4. Size distribution of the fragments digested by four combinations of REs in *Plutella xylostella*. (a) Digested *in silico* by *DDsilico*, (b) empirical digestion by *NlaIII* + *BfaI*, (c) empirical digestion by *NlaIII* + *AciI*, (d) empirical digestion by *NlaIII* + *EcoRI*, (e) empirical digestion by *NlaIII* + *MluCI*. The region between the dotted lines indicates the fragments between 300 bp and 450 bp.

## 4 Discussion

In our study, we simulated 131 species with two programs, and chose three species for experimental validation. The result showing linear relationship between the number of fragments and genome size is consistent with that of Yang *et al* (2016). It was also demonstrated that the digestion pattern of restriction enzymes among different species and insect orders was similar. We can choose the same combination to execute ddRAD protocol about insects researches. As for the combination of REs, *NlaIII* and *Taq*$^{\alpha}$*I* usually generated the highest number of predicted fragments with different overhangs. Such a large number of fragments may require deep sequencing to detect polymorphisms, given the potential for sequencing error. In insects, the cut sites for *NlaIII* and *Taq*$^{\alpha}$*I* are more frequent than for the other REs. Thus, *NlaIII* or *Taq*$^{\alpha}$*I* can provide a useful endonuclease combination for ddRAD studies. For the rare cut-sites enzyme, a six-base or four-base endonuclease could be used. Some studies have used four-base + six-base enzyme pairs, such as *EcoRI* + *MspI* in the original ddRAD protocol (Peterson *et al.*, 2012), and *PstI* + *MspI* for angiosperm plants (Yang *et al.*, 2016). However, these combinations may not be suitable for insects when large numbers of variants are needed, because the genome size of many insects is small, and there may be too few cut sites for six-base enzymes to generate sufficient SNPs. Combinations of four-base enzymes are commonly used for insects, e.g., *NlaIII* + *MluCI* in *Aedes aegypti* (Diptera, Culicidae) (Rašić *et al.*, 2014).

In ddRAD, digested fragments with different overhangs, generated by digestion of two REs, are selectively genotyped, while fragments with the same overhangs are excluded. Previous studies often used an experimental approach to evaluate the performance of potential combinations of REs for ddRAD. This approach will normally include fragments with the same overhangs as well as with different overhangs. Our study showed that there will be a high proportion of fragments with the same overhangs in most species, potentially biasing the selection of enzyme combinations when evaluated by an experimental approach. Simulation can help to avoid this problem, and it appears that the performance of different programs is similar for estimating the number of fragments. Among the four programs, *Digital_RADs.py* of BU-RAD-seq (https://github.com/BU-RAD-seq/Digital_RADs) (DaCosta & Sorenson, 2014), and DDRADSEQTOOLS (Mora-Márquez *et al.*, 2017) are computationally efficient, especially for large or complex genomes. *DDsilico* (Rašić *et al.*, 2014) running within Windows operating system can provide more information about digestion fragments, like the number of fragments with same or different overhangs, the total length of different range fragments. SimRAD (Lepais & Weir, 2014) is more suitable for phylogenetic ddRAD studies.

When choosing combinations of REs for ddRAD, three factors should be considered. The first factor is the species, like its lineage and its genome size. The performance of REs may differ among organisms and particular lineages. The second factor is the size range of target fragments. We considered fragments from 300 bp to 450 bp, and combinations of four-bases REs, such as *NlaIII* + *MluCI,* are more suitable for obtaining these or shorter fragments, as confirmed by the experimental data. Six-base enzymes should be used cautiously because of rare cut sites. When the number of rare cut sites is much smaller than common cut sites, numerous fragments with the same overhangs will be generated. The third factor is the compatibility of REs used in digestion. Some REs have special reaction characteristics. For example, *Taq*$^{\alpha}$*I* is sensitive to dam methylation, and *MluCI* generates the same overhang as for *EcoRI*. Care is needed when selecting suitable combinations of enzymes for ddRAD studies.

## 5 Conclusion

Our study revealed that the relative number of digested fragments based on different combinations of REs is conserved in many insects. The conserved profile of digestion will make it easier to determine the appropriate RE combination for ddRAD. Although our analyses were conducted on insects, the results may be relevant to other groups of organisms. When a large number of markers is needed for a ddRAD study on insects, combinations with *NlaIII* or *Taq*$^{\alpha}$*I* are recommended to generate fragments between 300 bp and 450 bp. Combinations with *EcoRI* or *MluCI* are recommended to allow for high sequencing depth with reduced sequencing costs.

**Funding** The research was funded by the Natural Science Foundation of Beijing Municipality (6162010), the National Natural Science Foundation of China (31472025), the International Cooperation Fund of Beijing Academy of Agriculture and Forestry Sciences (GJHZ2017), and the Beijing Key Laboratory of Environmentally Friendly Pest Management on Northern Fruits (BZ0432).

## References

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17: 81–92.

Andrews, K.R., Luikart, G. 2014. Recent novel approaches for population genomics data analysis. *Molecular Ecology*, 23: 1661–1667.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A, Selker, E.U., Cresko, W.A., Johnson, E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3: e3376.

Yin, C.L., Shen, G.Y., Guo, D.H., Wang, S.P., Ma, X.Z., Xiao, H.M., Liu, J.D., Zhang, Z., Liu, Y., Zhang, Y.Q., Yu, K.X., Huang, S.Q., Li, F. 2016. InsectBase: a resource for insect genomes and transcriptomes. *Nucleic Acids Research*, 44: D801–D807.

DaCosta, J.M., Sorenson, M.D. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PloS ONE*, 9: e106713.

DaCosta, J.M., Sorenson, M.D. 2015. ddRAD-seq phylogenetics based on nucleotide, indel, and presence-absence polymorphisms: Analyses of two avian genera with contrasting histories. *Molecular Phylogenetics and Evolution*, 94: 122–135.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., Blaxter, M.L. 2013. Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, 22: 3151–3164.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12: 499–510.

Derkarabetian, S., Burns, M., Starrett, J., Hedin, M. 2016. Population genomic evidence for multiple Pliocene refugia in a montane-restricted harvestman (Arachnida, Opiliones, Sclerobunus robustus) from the southwestern United States. *Molecular Ecology*, 25: 4611–4631.

Franchini, P., Monne Parera, D., Kautt, A.F., Meyer, A. 2017. quaddRAD: a new high-multiplexing and PCR duplicate removal ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage. *Molecular Ecology*, 26: 2783–2795.

Hoffberg, S.L., Kieran, T.J., Catchen, J.M., Devault, A., Faircloth, B.C., Mauricio, R., Glenn, T.C. 2016. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, 16: 1264–1278.

Johnson, J.S., Gaddis, K.D., Cairns, D.M., Konganti, K., Krutovsky, K.V. 2017. Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. *American Journal of Botany*, 104: 439–450.

Lepais, O., Weir, J.T. 2014. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, 14: 1314–1321.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos,K.E., Reed, L.K., Antolin, M.F., Storfer, A. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17: 142–152.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17: 240–248.

Mora-Márquez, F., Garcia-Olivares, V., Emerson, B.C., Lopez de Heredia, U. 2017. DDRADSEQTOOLS: a software package for in silico simulation and testing of double-digest RADseq experiments. *Molecular Ecology Resources*, 17: 230–246.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, 7: e37135.

Pukk, L., Ahmad, F., Hasan, S., Kisand, V., Gross, R., Vasemägi, A. 2015. Less is more: extreme genome complexity reduction with ddRAD using Ion Torrent semiconductor technology. *Molecular Ecology Resources*, 15: 1145–1152.

Puritz, J.B., Matz, M.V., Toonen, R.J., Weber, J.N., Bolnick, D,I., Bird, C.E. 2014. Demystifying the RAD fad. *Molecular Ecology*, 23: 5937–5942.

Rašić, G., Filipović, I., Weeks, A.R., Hoffmann, A.A. 2014. Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genomics*, 15: 275.

Rašić, G., Schama, R., Powell, R., Maciel-de Freitas, R., Endersby-Harshman, N.M., Filipović, I., Sylvestre, G., Máspero, R.C., Hoffmann, A.A. 2015. Contrasting genetic structure between mitochondrial and nuclear markers in the dengue fever mosquito from Rio de Janeiro: implications for vector control. *Evolutionary Applications*, 8: 901–915.

Russello, M.A., Waterhouse, M.D., Etter, P.D., Johnson, E.A. 2015. From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ*, 3: e1106.

Schweyen, H., Rozenberg, A., Leese, F. 2014. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological Bulletin*, 227: 146–160.

Tigano, A., Shultz, A.J., Edwards, S.V., Robertson, G.J., Friesen, V.L. 2017. Outlier analyses to test for local adaptation to breeding grounds in a migratory arctic seabird. *Ecology and Evolution*, 7: 2370–2381.

Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.L., Fernandez-Silva, I., Andrews, K.R., Bird, C.E. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, 1: e203.

Wang, S., Meyer, E., McKay, J.K., Matz, M.V. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, 9: 808–810.

Yang, G.Q., Chen, Y.M., Wang, J.P., Guo, C., Zhao, L., Wang, X.Y., Guo, Y., Li, L., Li, D.Z., Guo, Z.H. 2016. Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. *Plant Methods*, 12: 39.

**Table S1. List of the 131 insect genomes used for simulating analysis by *Digital_RADs.py* and *DDsilico*.**

| Order | Species name | GenBank accession number | Genome size (Mb) |
|---|---|---|---|
| Anoplura | *Pediculus humanus* | AAZO00000000 | 110.78 |
| Diplura | *Catajapyx aquilonaris* | JYFJ00000000 | 302.13 |
| Ephemeroptera | *Ephemera danica* | AYNC00000000 | 475.91 |
| Odonata | *Ladona fulva*[*] | APVN00000000 | 1158.11 |
| Orthoptera | *Locusta migratoria*[*] | AVCP00000000 | 5759.80 |
| Strepsiptera | *Mengenilla moldrzyki* | AGDA00000000 | 155.73 |
| Thysanoptera | *Frankliniella occidentalis* | JMDY00000000 | 415.78 |
| Trichoptera | *Limnephilus lunatus*[*] | JDSM00000000 | 1345.86 |
| Blattodea | *Blattella germanica*[*] | JPZV00000000 | 2037.20 |
| | *Zootermopsis nevadensis* | AUST00000000 | 485.03 |
| Coleoptera | *Agrilus planipennis* | JENH00000000 | 353.56 |
| | *Anoplophora glabripennis* | AQHT00000000 | 707.73 |
| | *Hypothenemus hampei* | LBGY00000000 | 151.27 |
| | *Leptinotarsa decemlineata* | AYNB00000000 | 678.27 |
| | *Onthophagus taurus* | JHOM00000000 | 270.53 |
| | *Priacma serrata*[**] | AGRH00000000 | 12.08 |
| | *Tribolium castaneum* | AAJJ00000000 | 165.94 |
| Lepidoptera | *Bombyx mori* | BABU00000000 | 431.70 |
| | *Chilo suppressalis* | ANCD00000000 | 314.17 |
| | *Heliconius melpomene* | CAFA00000000 | 269.66 |
| | *Melitaea cinxia* | APLT00000000 | 389.91 |
| | *Papilio glaucus* | JWHW00000000 | 374.82 |
| | *Papilio polytes* | BBJD00000000 | 227.02 |
| | *Papilio xuthus* | BBJE00000000 | 237.94 |
| | *Plutella xylostella* | AHIO00000000 | 336.85 |
| | *Spodoptera frugiperda* | JQCY00000000 | 514.23 |
| Hemiptera | *Acyrthosiphon pisum* | ABLF00000000 | 541.69 |
| | *Cimex lectularius* | JHUN00000000 | 513.61 |
| | *Dactylopius coccus*[**] | JMCM00000000 | 18.61 |
| | *Diaphorina citri* | AWGM00000000 | 485.71 |
| | *Gerris buenoi* | JHBY00000000 | 693.68 |
| | *Halyomorpha halys*[*] | JMPT00000000 | 1150.11 |
| | *Homalodisca vitripennis*[*] | JJNS00000000 | 2204.90 |
| | *Nilaparvata lugens*[*] | AOSB00000000 | 1140.79 |
| | *Oncopeltus fasciatus*[*] | JHQO00000000 | 1098.67 |
| | *Pachypsylla venusta* | AZLD00000000 | 701.76 |
| | *Piezodorus guildinii*[**] | JTEQ00000000 | 3.18 |
| | *Rhodnius prolixus* | ACPB00000000 | 564.63 |
| Hymenoptera | *Acromyrmex echinatior* | AEVX00000000 | 288.51 |
| | *Apis dorsata* | AUPE00000000 | 219.17 |
| | *Apis florea* | AEKZ00000000 | 230.49 |
| | *Apis mellifera* | JSUV00000000 | 229.11 |
| | *Athalia rosae* | AOFN00000000 | 156.83 |
| | *Atta cephalotes* | ADTU00000000 | 317.67 |
| | *Bombus impatiens* | AEQM00000000 | 244.29 |
| | *Bombus terrestris* | AELG00000000 | 248.65 |

**Table S1 (continued)**

| Order | Species name | GenBank accession number | Genome size (Mb) |
|---|---|---|---|
| | *Camponotus floridanus* | AEAB00000000 | 232.69 |
| | *Cephus cinctus* | AMWH00000000 | 162.25 |
| | *Cerapachys biroi* | JASI00000000 | 212.83 |
| | *Copidosoma floridanum* | JBOX00000000 | 555.05 |
| | *Cotesia vestalis* | JZSA00000000 | 186.10 |
| | *Fopius arisanus* | JRKH00000000 | 153.63 |
| | *Harpegnathos saltator* | AEAC00000000 | 294.47 |
| | *Linepithema humile* | ADOQ00000000 | 219.50 |
| | *Megachile rotundata* | AFJA00000000 | 272.66 |
| | *Microplitis demolitor* | AZMT00000000 | 241.19 |
| | *Monomorium pharaonis* | BBSX00000000 | 257.98 |
| | *Nasonia giraulti* | ADAO00000000 | 283.61 |
| | *Nasonia longicornis* | ADAP00000000 | 285.73 |
| | *Nasonia vitripennis* | AAZX00000000 | 295.78 |
| | *Orussus abietinus* | AZGP00000000 | 201.22 |
| | *Pogonomyrmex barbatus* | ADIH00000000 | 235.65 |
| | *Solenopsis invicta* | AEAQ00000000 | 396.03 |
| | *Trichogramma pretiosum* | JARR00000000 | 195.09 |
| | *Vollenhovia emeryi* | BBUO00000000 | 287.90 |
| | *Wasmannia auropunctata* | BBSV00000000 | 324.12 |
| Diptera | *Aedes aegypti*[*] | AAGE00000000 | 1278.73 |
| | *Anopheles albimanus* | APCK00000000 | 173.34 |
| | *Anopheles arabiensis* | APCN00000000 | 246.57 |
| | *Anopheles atroparvus* | AXCP00000000 | 224.29 |
| | *Anopheles christyi* | APCM00000000 | 172.66 |
| | *Anopheles coluzzii* | ABKP00000000 | 224.42 |
| | *Anopheles culicifacies* | AXCM00000000 | 203.00 |
| | *Anopheles darlingi* | ADMH00000000 | 136.94 |
| | *Anopheles dirus* | APCL00000000 | 216.31 |
| | *Anopheles epiroticus* | APCJ00000000 | 223.49 |
| | *Anopheles farauti* | JXXC00000000 | 175.82 |
| | *Anopheles funestus* | APCI00000000 | 225.22 |
| | *Anopheles gambiae* | ABKQ00000000 | 454.69 |
| | *Anopheles koliensis* | JXXB00000000 | 151.11 |
| | *Anopheles maculatus* | AXCL00000000 | 418.51 |
| | *Anopheles melas* | AXCO00000000 | 224.16 |
| | *Anopheles merus* | AXCQ00000000 | 288.05 |
| | *Anopheles minimus* | APHL00000000 | 201.79 |
| | *Anopheles nili* | ATLZ00000000 | 98.32 |
| | *Anopheles punctulatus* | JXXA00000000 | 146.16 |
| | *Anopheles quadriannulatus* | APCH00000000 | 283.83 |
| | *Anopheles sinensis* | ATLV00000000 | 214.51 |
| | *Anopheles stephensi* | ALPR00000000 | 209.48 |
| | *Bactrocera cucurbitae* | JRNW00000000 | 374.82 |
| | *Bactrocera dorsalis* | JFBF00000000 | 414.99 |
| | *Bactrocera tryoni* | JHQJ00000000 | 519.01 |

**Table S1 (continued)**

| Order | Species name | GenBank accession number | Genome size (Mb) |
|---|---|---|---|
| | *Belgica antarctica* | JPYR00000000 | 89.58 |
| | *Ceratitis capitata* | AOHK00000000 | 479.05 |
| | *Chironomus tentans* | CBTT00000000 | 213.46 |
| | *Culex quinquefasciatus* | AAWU00000000 | 579.04 |
| | *Drosophila albomicans* | ACVV00000000 | 251.21 |
| | *Drosophila ananassae* | AAPP00000000 | 230.99 |
| | *Drosophila biarmipes* | AFFD00000000 | 168.60 |
| | *Drosophila bipectinata* | AFFE00000000 | 166.41 |
| | *Drosophila elegans* | AFFF00000000 | 171.27 |
| | *Drosophila erecta* | AAPQ00000000 | 152.71 |
| | *Drosophila eugracilis* | AFPQ00000000 | 156.94 |
| | *Drosophila ficusphila* | AFFG00000000 | 152.44 |
| | *Drosophila grimshawi* | AAPT00000000 | 200.47 |
| | *Drosophila kikkawai* | AFFH00000000 | 164.29 |
| | *Drosophila melanogaster* | JSAE00000000 | 143.73 |
| | *Drosophila miranda* | AJMI00000000 | 132.59 |
| | *Drosophila mojavensis* | AAPU00000000 | 193.83 |
| | *Drosophila persimilis* | AAIZ00000000 | 188.37 |
| | *Drosophila rhopaloa* | AFPP00000000 | 197.38 |
| | *Drosophila sechellia* | JAQR00000000 | 166.59 |
| | *Drosophila simulans* | JPYS00000000 | 124.97 |
| | *Drosophila suzukii* | AWUT00000000 | 202.30 |
| | *Drosophila takahashii* | AFFI00000000 | 181.03 |
| | *Drosophila virilis* | AANI00000000 | 206.03 |
| | *Drosophila willistoni* | AAQB00000000 | 235.52 |
| | *Drosophila yakuba* | AAEU00000000 | 165.71 |
| | *Glossina austeni* | JMRR00000000 | 370.27 |
| | *Glossina brevipalpis* | JFJS00000000 | 315.36 |
| | *Glossina fuscipes* | JFJR00000000 | 374.78 |
| | *Glossina morsitans* | JXPS00000000 | 355.59 |
| | *Glossina pallidipes* | JMRQ00000000 | 357.33 |
| | *Glossina palpalis* | JXJN00000000 | 380.10 |
| | *Lucilia cuprina* | JHUJ00000000 | 434.13 |
| | *Lutzomyia longipalpis* | AJWK00000000 | 154.23 |
| | *Mayetiola destructor* | AEGA00000000 | 152.72 |
| | *Megaselia scalaris* | CAQQ00000000 | 489.35 |
| | *Musca domestica* | AQPM00000000 | 691.74 |
| | *Phlebotomus papatasi* | AJVK00000000 | 489.35 |
| | *Stomoxys calcitrans*[*] | LDNW00000000 | 1086.14 |

[*]Insect species with genome size more than 1 Gb;
[**]Genome sequences with less than 20 Mb.

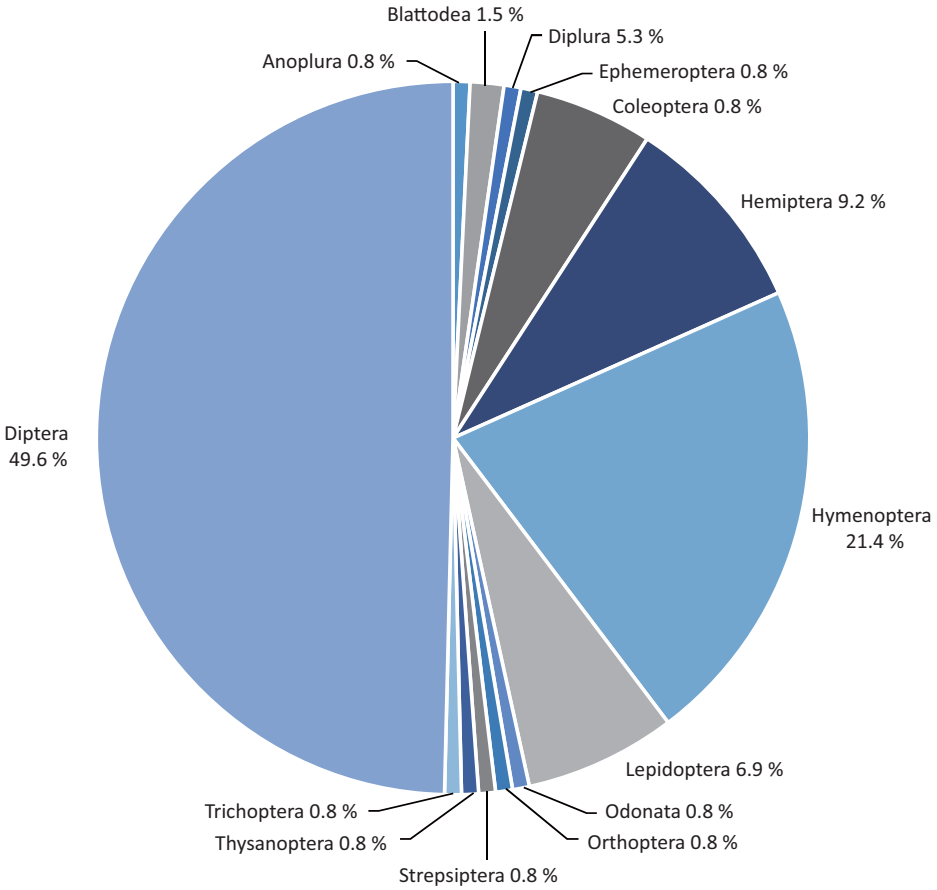


Figure S1. Distribution of the 131 genomes across 15 insect orders. Data were downloaded from InsectBase.

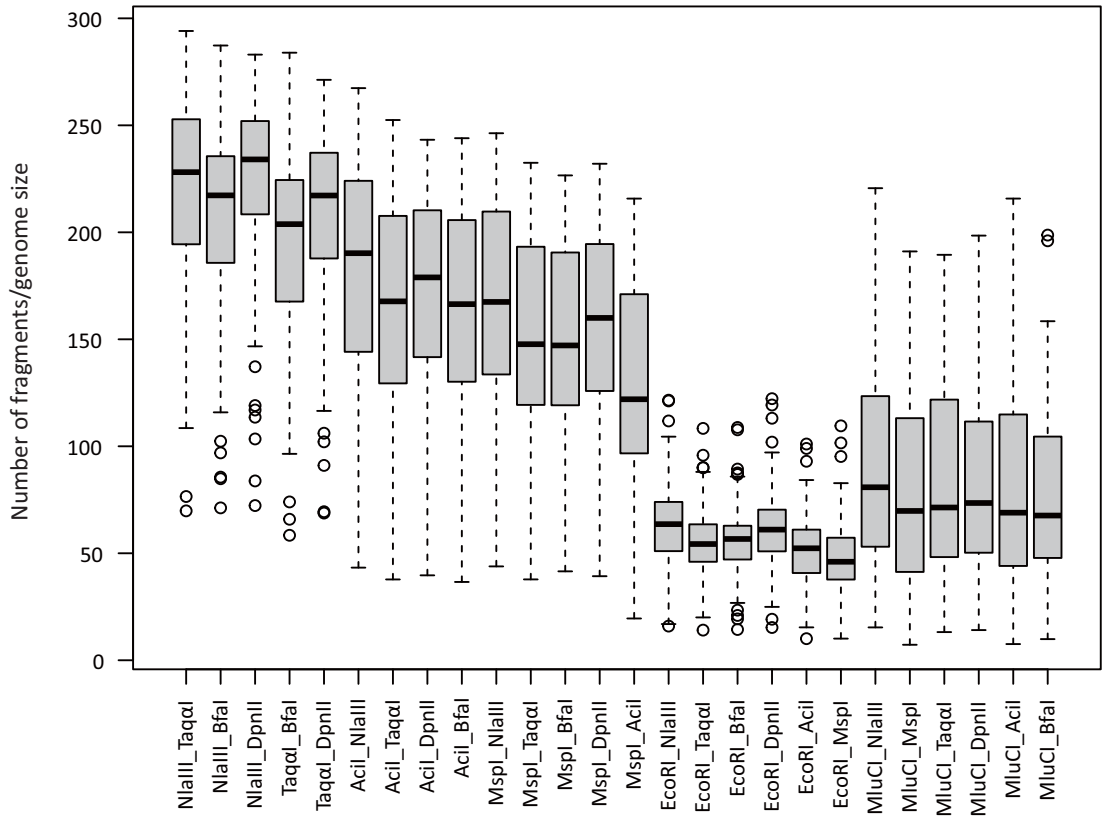


Figure S2. Boxplots of the number of fragments (300–450 bp) with different overhangs digested by 26 combinations of restriction endonucleases in 131 insect genomes simulated by the *Digital_RADs.py* program. The x-axis shows different combinations of restriction endonuclease pairs. Circles indicate outliers.

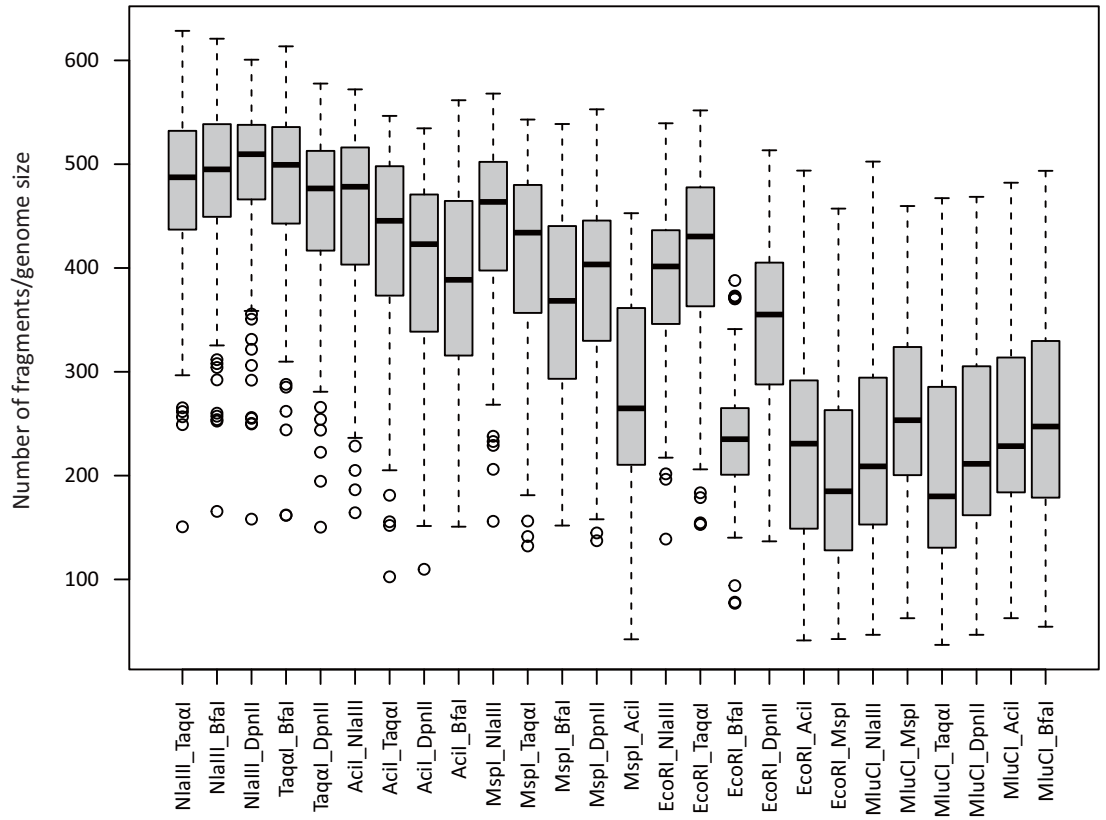


Figure S3. Boxplots for the number of total fragments (300-450 bp) digested by 26 combinations of restriction endonucleases in 131 insect genomes simulated by the *DDsilico* program. The x-axis shows different combinations of restriction endonuclease pairs. Circles indicate outliers.

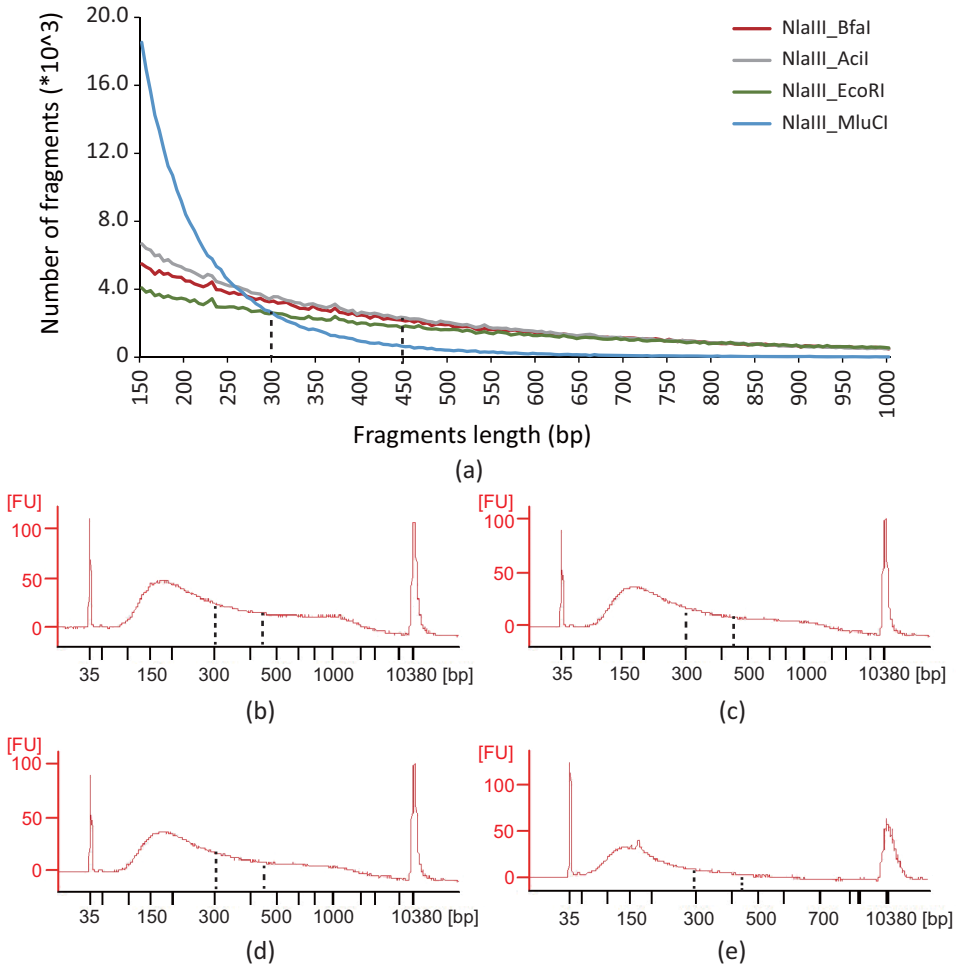


Figure S4. Size distribution of the fragments digested by four combinations of REs in *Apis mellifera*. (a) Digested in silico by *DDsilico*, (b) empirical digestion by *NlaIII* + *BfaI*, (c) empirical digestion by *NlaIII* + *AciI*, (d) empirical digestion by *NlaIII* + *EcoRI*, (e) empirical digestion by *NlaIII* + *MluCI*. The region between the dotted lines indicates the fragments between 300 bp and 450 bp.

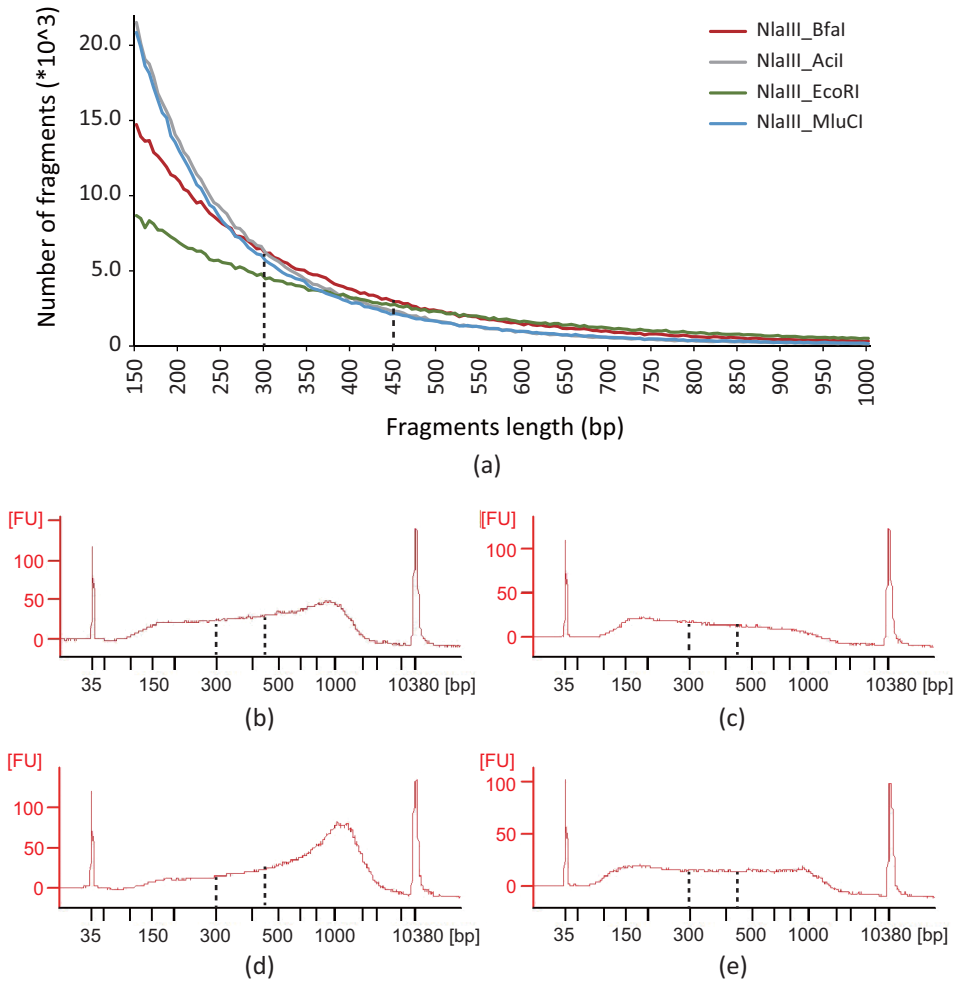


Figure S5. Size distribution of the fragments digested by four combinations of REs in *Frankliniella occidentalis*. (a) digested in silico by *DDsilico*, (b) empirical digestion by *NlaIII* + *BfaI*, (c) empirical digestion by *NlaIII* + *AciI*, (d) empirical digestion by *NlaIII* + *EcoRI*, (e) empirical digestion by *NlaIII* + *MluCI*. The region between the dotted lines indicates the fragments between 300 bp and 450 bp.