Supplementary material for the manuscript: "The worldwide invasion of *Drosophila suzukii* is accompanied by a large increase of transposable element load and a small number of putatively adaptive insertions "

Authors:

Vincent Mérel [1], Patricia Gibert [1], Inessa Buch [1], Valentina Rodriguez Rada [1], Arnaud Estoup [2], Mathieu Gautier [2], Marie Fablet [1], Matthieu Boulesteix [1*], Cristina Vieira [1*]

*** co-corresponding authors:**

matthieu.boulesteix@univ-lyon1.fr

cristina.vieira@univ-lyon1.fr

Affiliations:

[1] Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

[2] CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

Key words: *Drosophila suzukii*, Transposable Elements, Biological Invasion, Populations, Adaptation, PoolSeq.

Supplementary figures

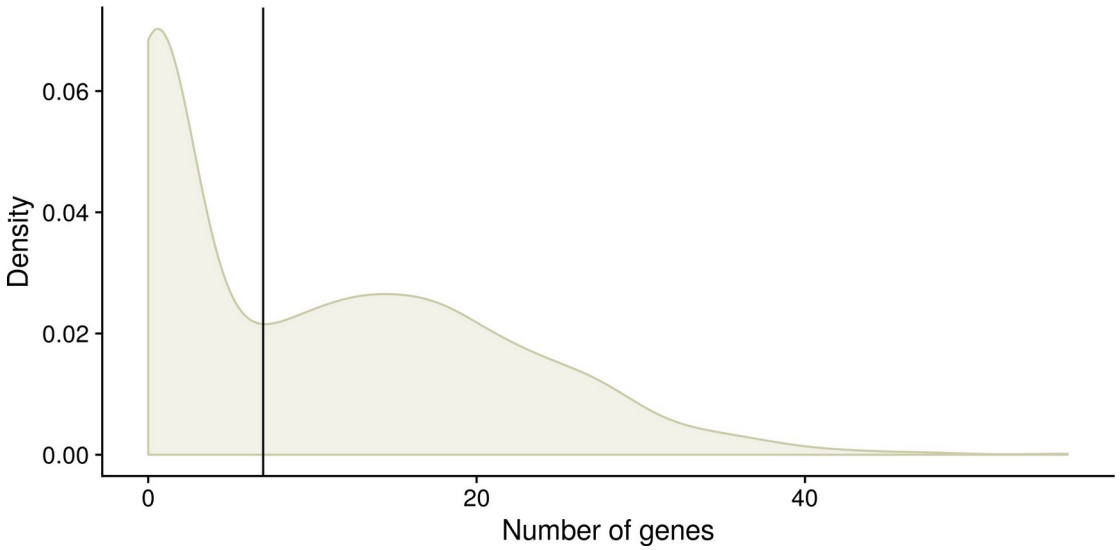

Figure 1:
Distribution of the number of genes per 200 kb windows in *D. suzukii* assembly. The vertical line corresponds to x=7.

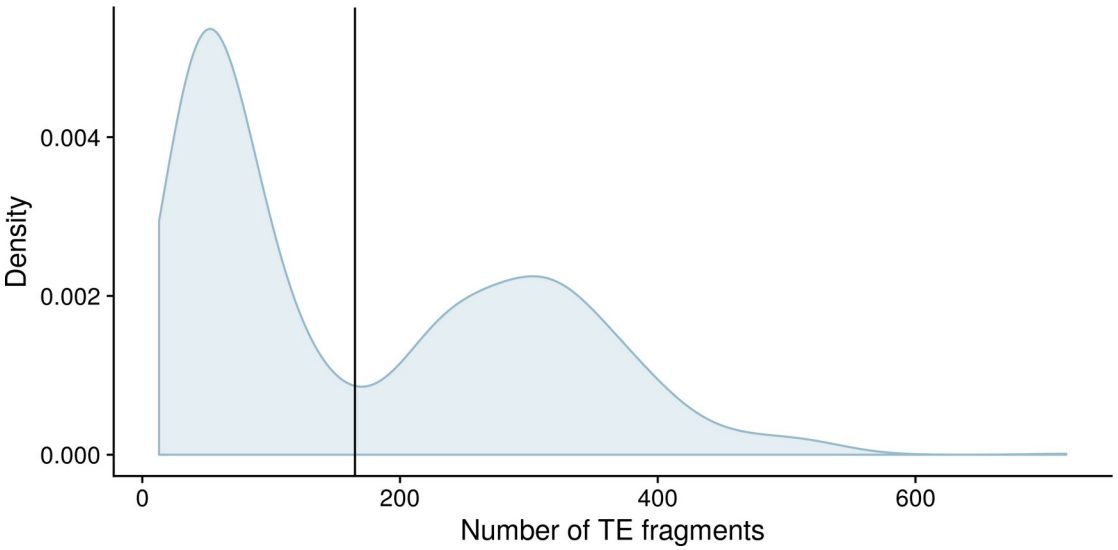

Figure 2: Distribution of the number of TE fragments per 200 kb windows in *D. suzukii* assembly. The vertical line corresponds to x=165.

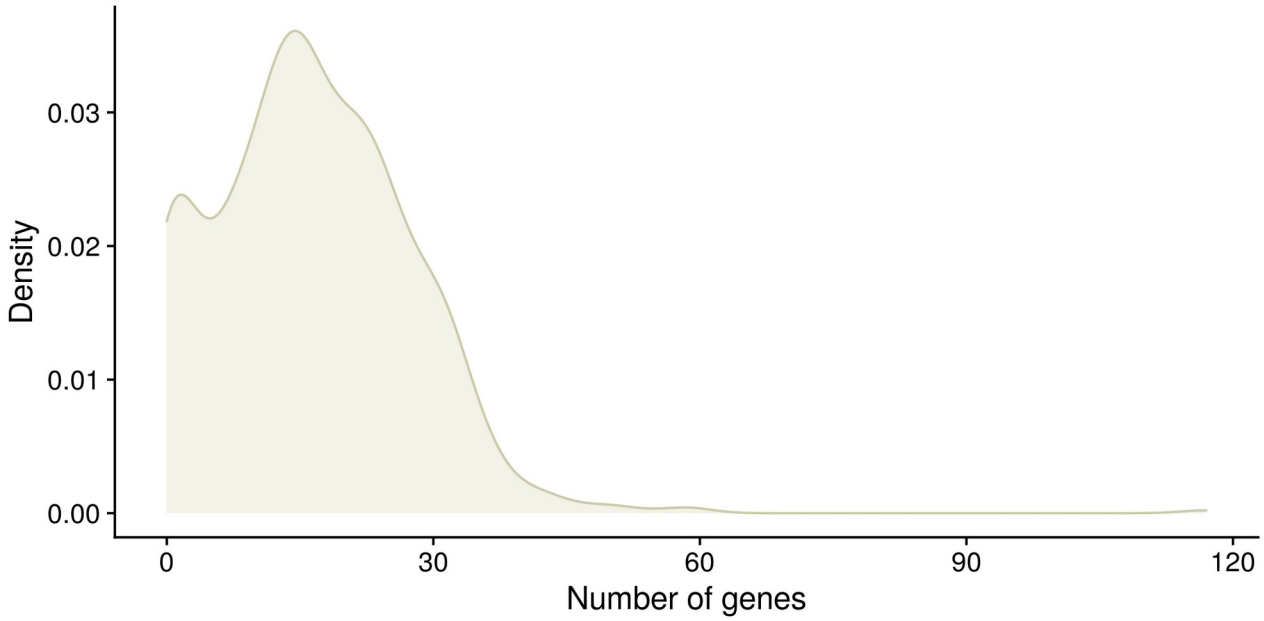

Figure 3:

Distribution of the number of genes per 200 kb windows in *D. melanogaster* assembly.

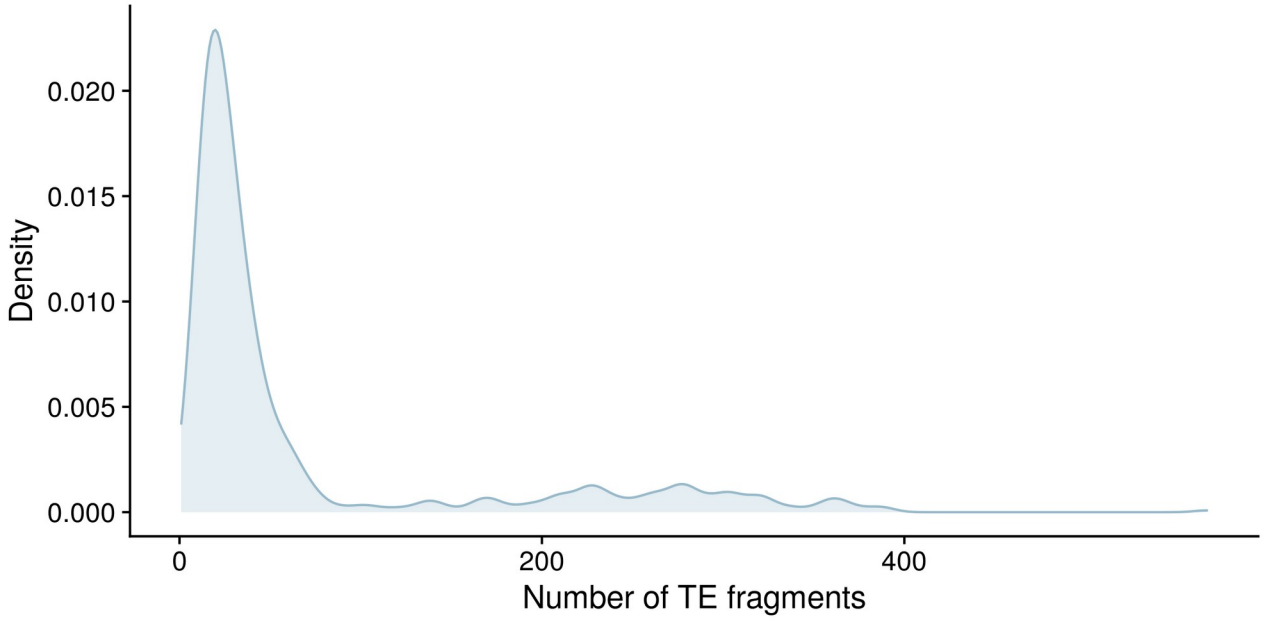

Figure 4:

Distribution of the number of TE fragments per 200 kb windows in *D. melanogaster* assembly.

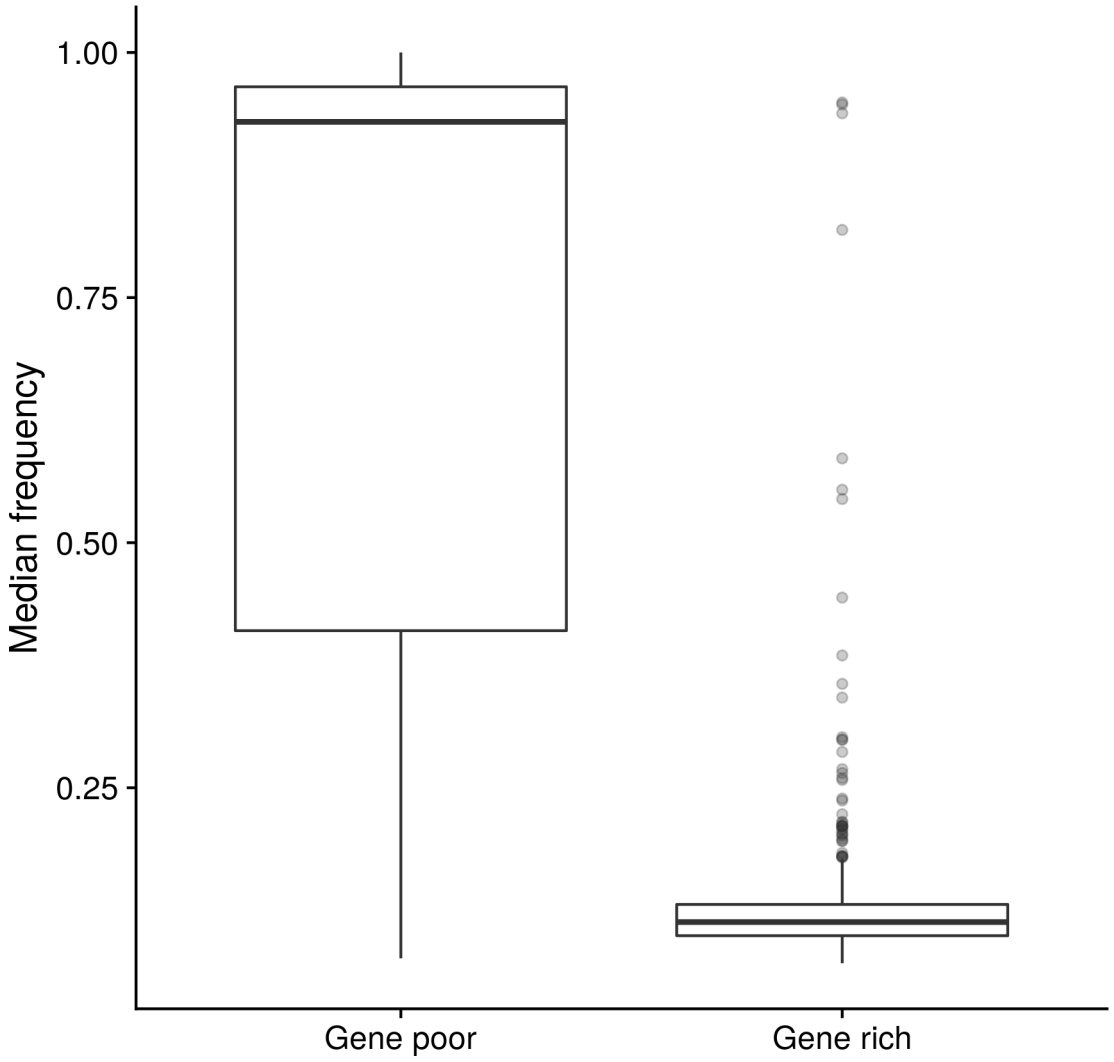

Figure 5:

Distribution of the median TE insertion frequency per 200 kb windows in *D. suzukii* assembly for gene-poor (< 7 genes per Mb) or gene-rich (≥ 7 genes per Mb) windows. Frequencies were estimated in the Watsonville reference population.

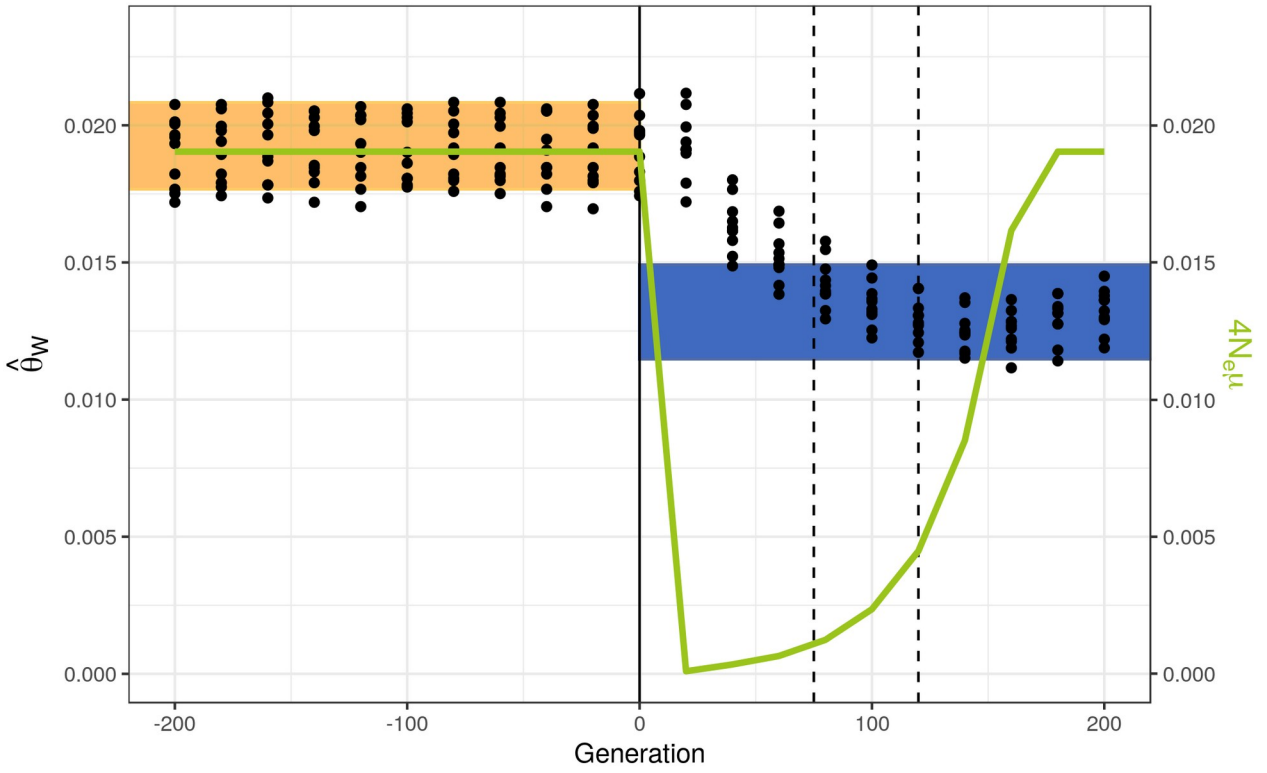

Figure 6:

Evolution of $\hat{\theta}_W$ in a simulated population undergoing a bottleneck. Mutation and recombination rates come from *D. melanogaster* studies (see Materials & Methods). The initial population size mimics the expected population size in native populations. At generation 0, population size is divided by 200, and then population size is multiplied by 1.9 every generation. The vertical solid line shows the bottleneck. For comparison, in our PoolSeq dataset sampling was done 75-120 generations after the bottleneck (considering a bottleneck occurring in 2008, a sampling between 2013 and 2015, and 15 generation per year). The dashed lines define the sampling period. The green line represents $4N_e\mu$, *i.e.* the expected value of $\hat{\theta}_W$ at the equilibrium. The orange rectangle represents the range of $\hat{\theta}_W$ in native populations. The blue rectangle represents the range of $\hat{\theta}_W$ in invasive populations.

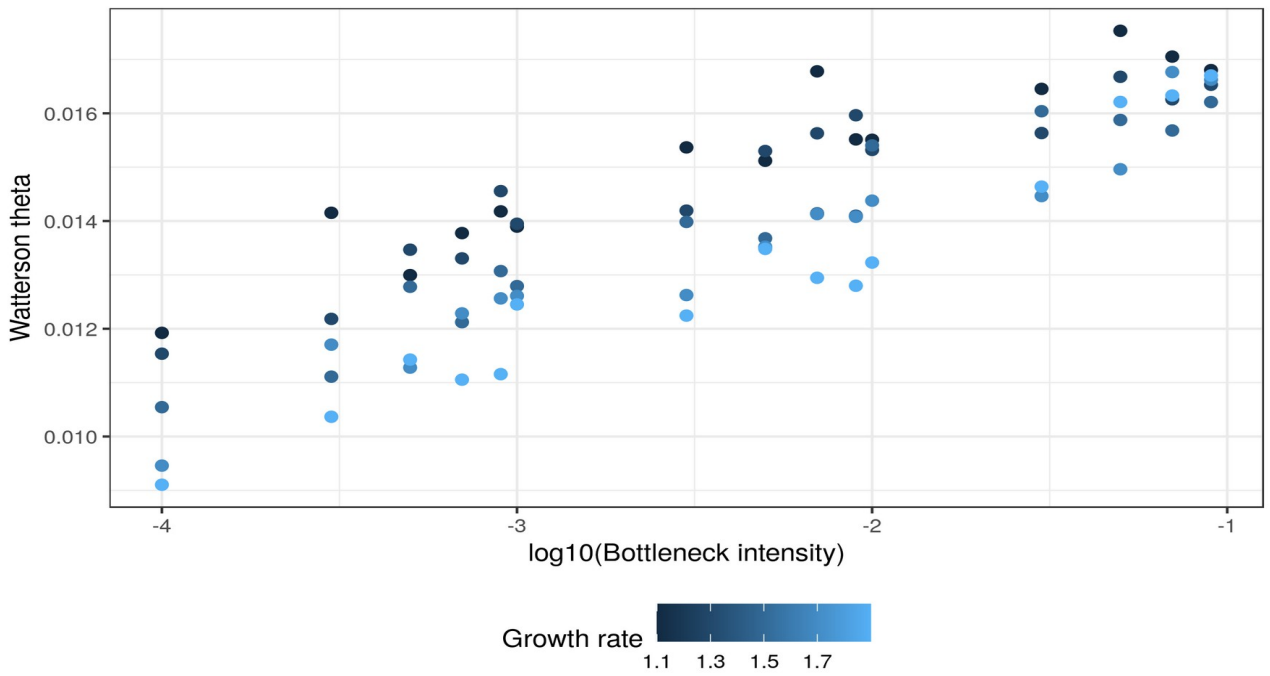

Figure 7:

Correlation between $\hat{\theta}_W$ and bottleneck intensity in simulated populations undergoing a bottleneck. Mutation and recombination rates come from *D. melanogaster* studies (see Materials & Methods). The initial population size mimics the expected population size in native populations. After the 7.5 N generations necessary to reach an equilibrium, population size is multiplied by a factor ranging from 0.0001 to 0.09, the bottleneck intensity. Then population size is then multiplied by a factor ranging from 2 to 2 every generation, the growth rate. $\hat{\theta}_W$ is calculated after 100 generations. For comparison, in our PoolSeq dataset sampling was done 75-120 generations after the bottleneck (considering a bottleneck occurring in 2008, a sampling between 2013 and 2015, and 15 generation per year). The value of $\hat{\theta}_W$ corresponds to the average over 10 replicates.

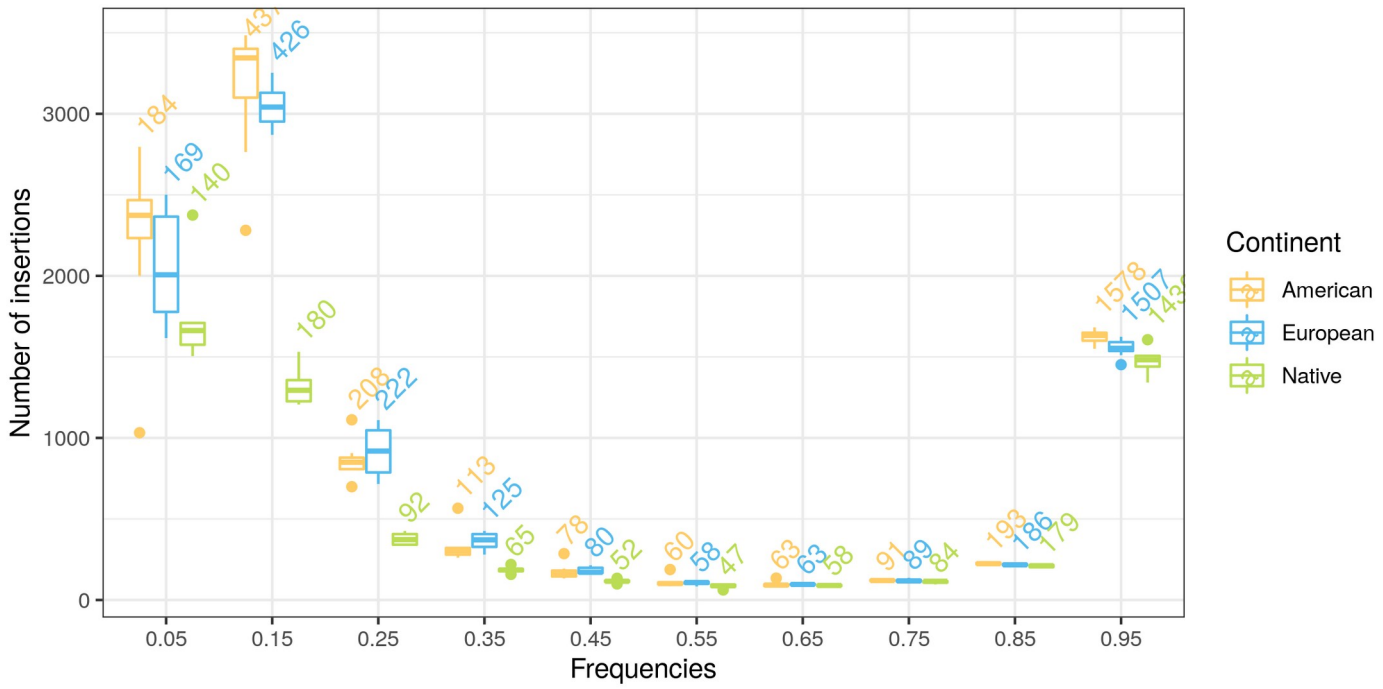

Figure 8:

Number of insertions in different categories of frequencies in *D. suzukii* populations. The numbers above boxes correspond to the mean number of insertions per haploid genome and colors illustrate continents.

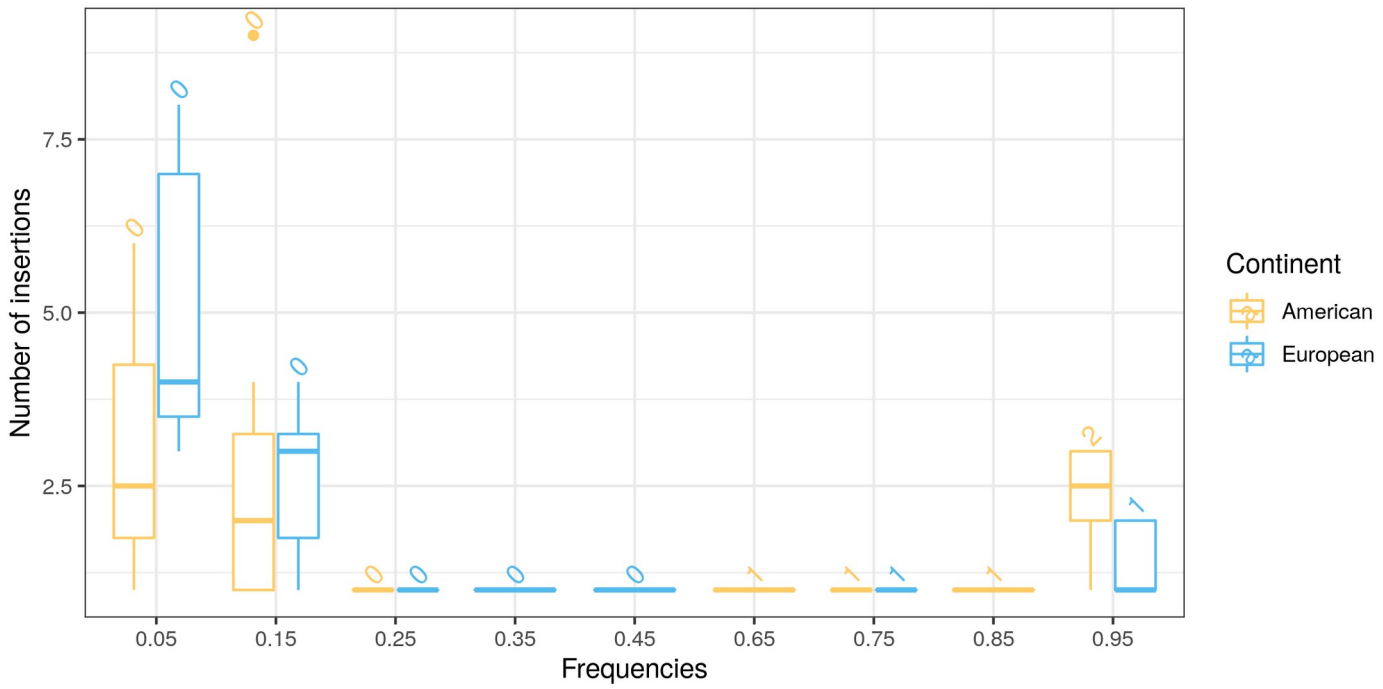

Figure 9:
Number of insertions in different categories of frequencies in *D. suzukii* populations for putatively horizontally transferred TEs. The numbers above boxes correspond to the mean number of insertions per haploid genome and colors illustrate continents.

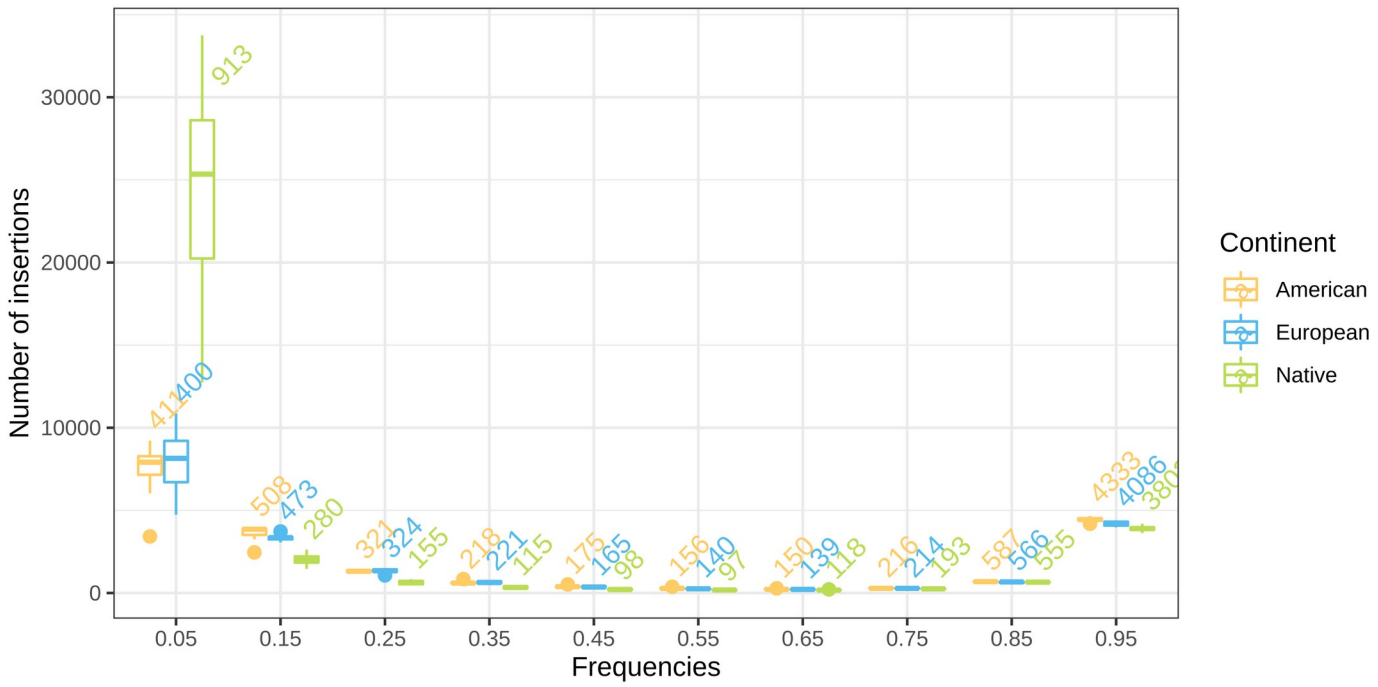

Figure 10:
Number of insertions in different categories of frequencies in *D. suzukii* populations (TE calling performed using the separate mode of PoPoolationTE2 and without reads subsampling). The numbers above boxes correspond to the mean number of insertions per haploid genome and colors illustrate continents.

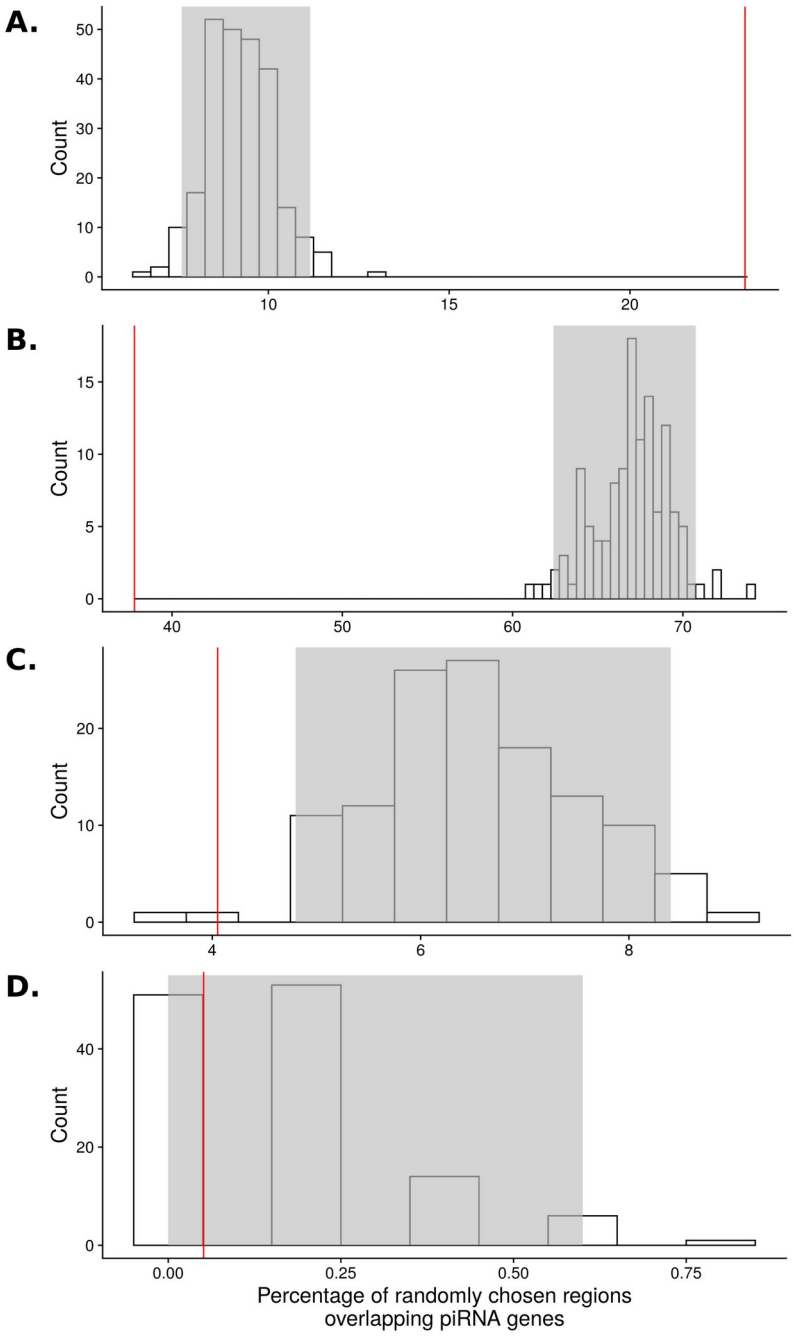

Figure 11:
Distribution of the percentage of randomly chosen regions surrounding SNPs in *D. suzukii* assembly and overlapping: A. repeated sequences; B. genes; C. genes encoding transcription factors (Tfs); D. genes of the piRNA pathway. 250 samples of 1000 regions were used to draw the distribution A., 125 samples of 500 regions for distributions B,C and D. The gray rectangle in the background delimits the portion of the distribution between quantile 2.5% and quantile 97.5%. The vertical red lines correspond to the observed percentage for regions associated with TE abundance.

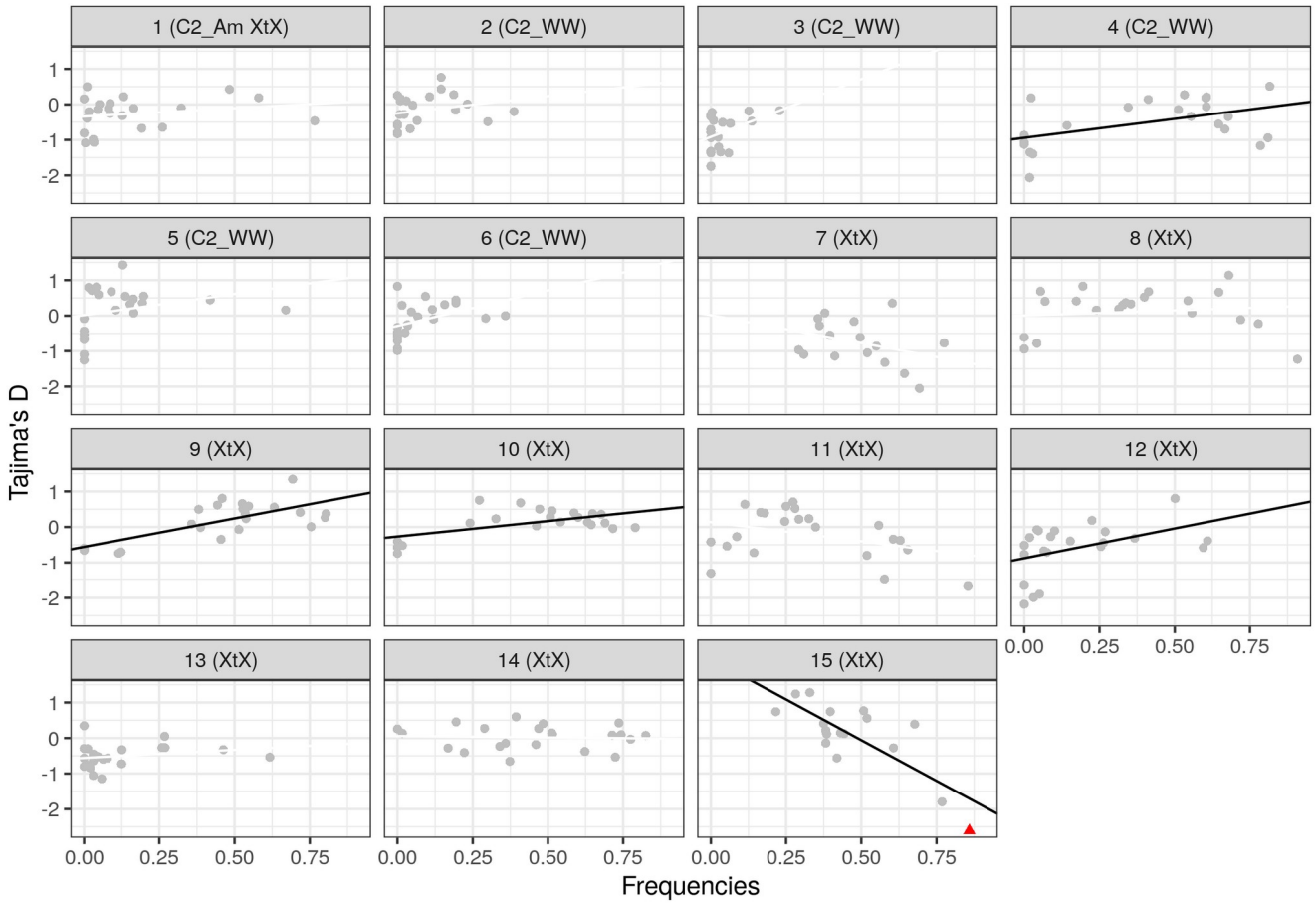

Figure 12:
Correlation between insertion frequencies and local Tajima's D estimates in the 22 *D. suzukii* populations for each of the 15 putatively adaptive insertions. Each panel corresponds to one insertion and Tajima's D are estimated from the 1 kb window containing the insertion. Regression lines are drawn when linear correlations are significant (Pearson's product-moment correlation, $p<0.05$). The red dot indicates that local Tajima's D is inferior to quantile 5% in that population.

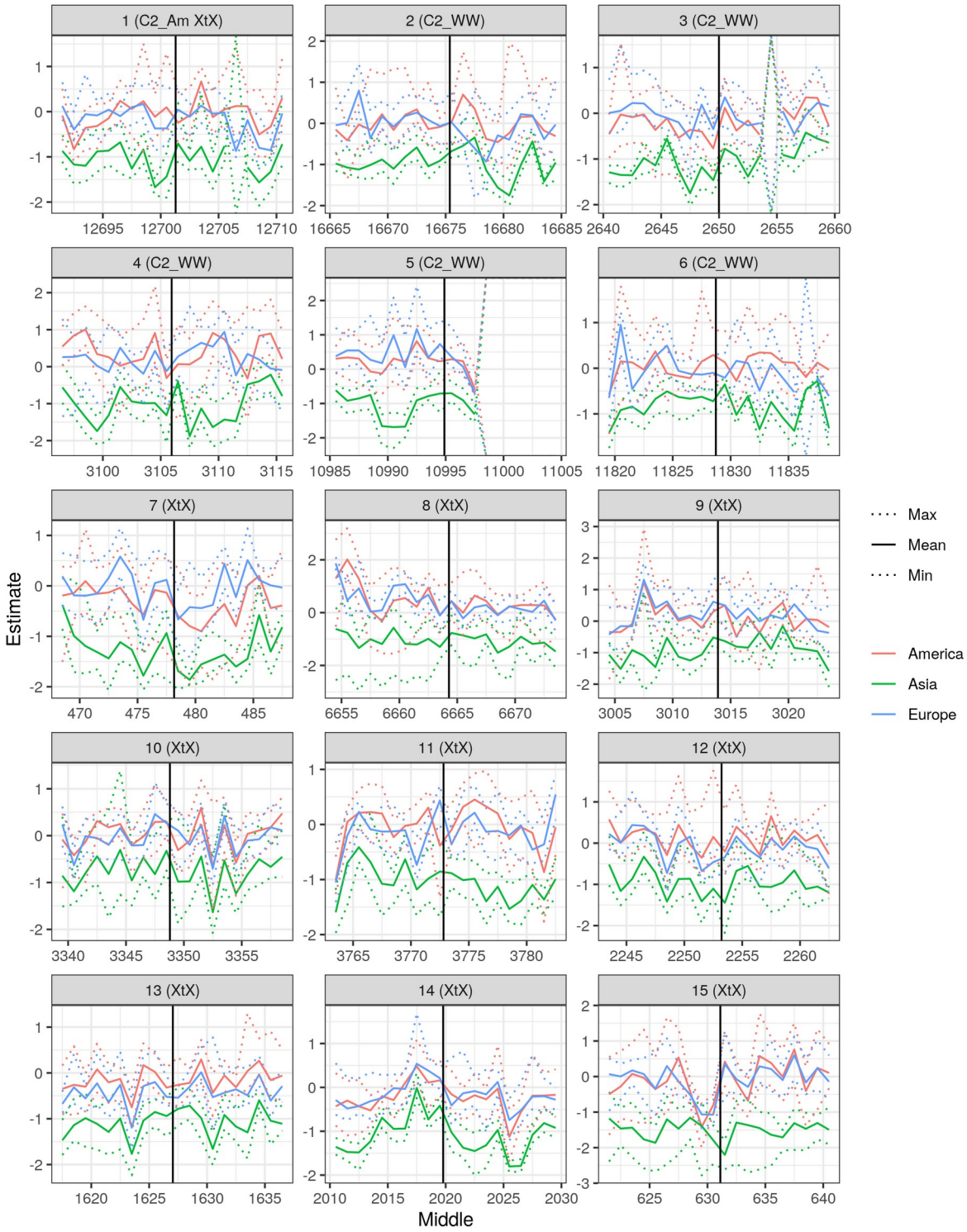

Figure 13:
Tajima's D around the 15 putatively adaptative insertions. Positions along the contigs (bp) are on the x axis and TE insertions are located at the vertical black lines. Each statistics is estimated using SNPs/InDels in a 1-kb genomic window. Asian populations are in green, American in red and European in blue.

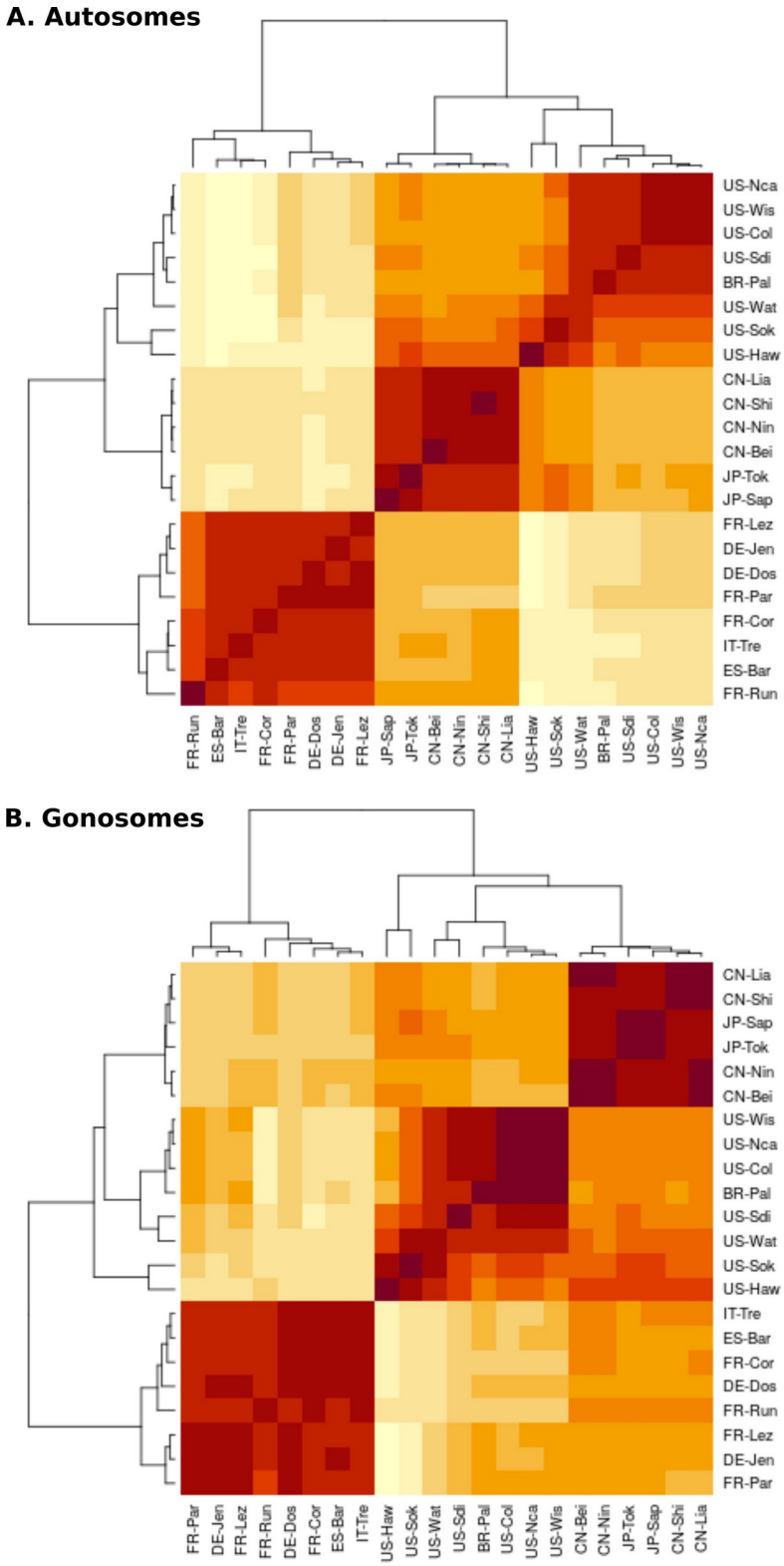

Figure 14:
Correlation plots of the scaled covariance matrices of population allele frequencies (Omega) among all 22 *D. suzukii* populations based on autosomal (A) and gonosomal (B) TE insertions.

Supplementary tables

Table 1:

Percentage of *D. suzukii* assembly occupied by each TE superfamily

CMC	0.40
Copia	0.23
CR1	2.41
DNA	0.05
Gypsy	13.65
hAT	0.19
hAT?	0.01
Helitron	6.95
I	2.48
Kolobok	0.05
L2	0.48
Maverick	4.92
Merlin	0.03
MULE	0.01
P	0.03
Pao	6.44
Penelope	0.00
PIF	0.40
PiggyBac	0.03
R1	3.24
R2	0.03
RTE	0.13
Sola	0.00
TcMar	0.87
Unknown	4.07
Zator	0.00

Table 2:

Number of Mb of *D. suzukii* assembly attributed to each of the *D. melanogaster* chromosomes

D. melanogaster chromosome	Mb of D. suzukii assembly
2L	51.9
2R	58.8
3L	45.6
3R	50.0
4	2.6
X	31.7

Table 3:

Number of families with median of High (f ≥0.75), Intermediate (0.25≤f<0.75), or Low (f<0.25) frequency in *D. suzukii* reference population for each TE order. Only families with more than 10 insertions are considered.

	DNA	LINE	LTR	RC	Unknown
High f.	6	9	13	1	6
Intermediate f.	1	2	0	0	1
Low f.	18	21	19	5	17

Supplementary methods

Creation of a TE database

A TE database was created by merging previously established consensus of Drosophila TE families and *de novo* reconstructed consensus of *D. suzukii* TE families. The previously established consensus were obtained by extracting all Drosophila consensus annotated as DNA, LINE, LTR, Other, RC, SINE and Unknown from Dfam and Repbase databases (release 2016-2018 for both) (Hubley et al. 2016; https://www.girinst.org/repbase/). Full LTR element sequences were reconstructed by merging LTRs and their internal parts. *De novo* reconstruction was performed using an assembly of an American strain from Watsonville, sequenced using PacBio long reads technology, and the REPET package (v2.5) (Flutre et al. 2011; Paris et al. 2020). Unless otherwise specified, the options were used as in the default configuration file. Briefly, the genome assembly was cut into batches and aligned to itself using blastn (ncbi-blast v2.2.6) (Altschul et al. 1990). High-scoring Segment Pairs (HSPs) were clustered using Recon (v1.08) and Piler (v1.0) (Bao and Eddy 2002; Edgar and Myers 2005). A structural detection step was performed using LTRHarvest from the GenomeTools package (v1.5.8) (Ellinghaus et al. 2008; Gremme et al. 2013). LTRHarvest-produced sequences were clustered using blastclust. Consensus sequences were created for each cluster using MAP (Huang 1994). Additional consensus sequences were generated using RepeatScout (v1.0.5) (Price et al. 2005). All consensus, *i.e.* from Recon, Piler, LTRHarvest and RepeatScout, were further submitted to a filtering step. Sequences were retained only if they produced at least 3 hits against the genome assembly with at least 98% query coverage (blastn, blast 2.6.0+). Structural and coding features were identified and used to classify consensus (see Hoede et al.(2014) for classification details, the used libraries were ProfilesBankForREPET_Pfam27.0_GypsyDB.hmm, repbase20.05_aaSeq_cleaned_TE.fsa, repbase20.05_ntSeq_cleaned_TE.fsa). Single satellite repeats, potential host genes and unclassified sequences were filtered out. Since REPET can easily mis-annotate any pair of repeats separated by a spacer as TRIM or LARD, those sequences were also removed (Arkhipova 2017). Remaining sequences were further annotated by homology to previously established consensus of Drosophila TE families. Homology was determined using RepeatMasker (-cutoff 250, v 1.332) (http://www.repeatmasker.org/). We followed the rules below: 1) if all hits belonged to the same superfamily, the sequence was annotated as corresponding to that particular superfamily and order; 2) if hits from different superfamilies were observed the sequence was considered as ambiguous; 3) without any hit, the sequence was annotated as unknown. Ambiguous sequences were manually curated, sequences which could be unambiguously attributed to one superfamily according to hits and proteic domains were kept (proteic domains were investigated using NCBI Conserved Domain Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Finally, consensus were clustered in families using UClust (-id 0.80, -strand both, –maxaccepts 0 –maxrejects 0; v11.0.667) (Edgar 2010). The annotation, superfamily and order, attributed to each cluster, *i.e.* each family, is the annotation of the longest sequence in the cluster. The generated TE database is accessible at: https://github.com/vmerel/Dsu-TE.

Investigation of Watterson's theta evolution

In order to determine whether the observed values of $\hat{\theta}_w$ in the invasive populations could be be the consequence of a simple bottleneck from native populations we performed forward simulations using SLiM (v3.5) (Haller and Messer 2019). Using a chromosome of length 1 kb, a recombination rate of $2.32 \times 10^{-}$

[8] (Comeron et al. 2012) and a mutation rate of 2.8x10^{-9} (Keightley et al. 2014), we followed $\hat{\theta}_w$ in a population that undergo a bottleneck. The initial population size was chosen to equal our estimation of native populations size. Assuming that Watterson's theta estimator did achieve equilibrium in these populations, we can use $\hat{\theta}_w = 4 N_e * \mu$ to estimate the population size. Given a mean $\hat{\theta}_w$ of 0.019 in these populations and a mutation rate of 2.8x10^{-9} (Keightley et al. 2014), population size should be 1.7x10^{7} individuals. After a burnin period of $7.5 N_e$ the population size was divided by a factor ranging from 11 to 10,000. This bottleneck was followed by a period of population expansion with an exponential growth rate comprised between 1 and 2. Ten replicates were performed by combination of factors. Note that to improve computing time a 0.05 downscaling was performed.

Appendix A: Validation of the TE calling

The accuracy of the TE calling procedure was validated by a simulation work using simulaTE (Kofler 2018). As a starting point, an artificial genome devoid of TEs was created by removing masked nucleotides in a randomly selected 1 Mb chunk of the masked assembly. The resulting genome was 607,100 bp long. A population of 1000 diploid individuals each displaying 500 insertions, of frequencies ranging from 0.01 to 0.99, was generated by inserting TE sequences in the artificial genome. A reference genome, containing 250 of these insertions was also created. This artificial population was used to simulate read data corresponding to each of the 22 PoolSeq samples. For each sample, x haploid genomes were drawn according to the exact number of individuals in the sample. Reads were simulated using simulaTE and mimicking the coverage and insert size of the original sample. We used the coverage estimated in Olazcuaga et al. (2020) and the inner distance, i.e. insert size -2*Read length extracted from the ppileup file header. The standard deviation on the inner distance was set to 100 bp. The TE frequency pipeline, (joint mode of PoPoolationTE2), and TE abundance pipeline (separate mode of PoPoolationTE2 and with reads subsampling), described in the Materials and Methods section were then run on this dataset

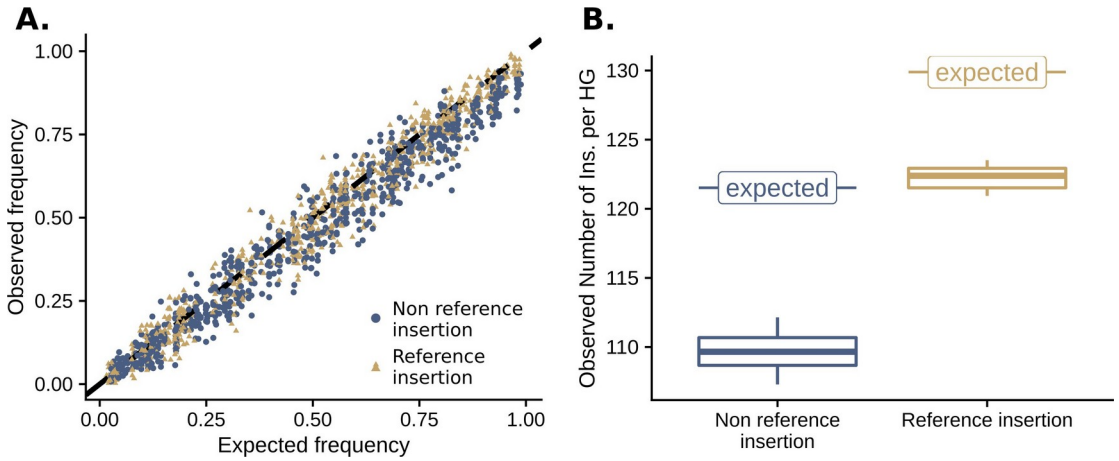

Figure a1:
Validation of the TE frequency and TE abundance pipelines: expectations vs observations in a 22 samples simulated dataset mimicking the original dataset A. Estimation of TE insertion frequencies. Observed frequencies in the simulated dataset are compared to expected frequencies. Insertions absent from the reference genome are shown in blue whereas insertions present in the reference genome are in gold. **B. Estimation of TE abundance**. Distribution of the numbers of haploid insertions for the 22 simulated samples for both non reference insertions and reference insertions. The expected numbers of insertions per haploid genome (HG), 121.5 for non reference insertions and 129.9 for reference insertions, are indicated by a horizontal segment.

A run of our pipelines on a simulated dataset mimicking the original *D. suzukii* dataset indicated that our methods estimate accurately TE insertion frequencies and TE abundances (as the numbers of TE insertions

per haploid genome (HG) per population). Regarding the TE frequency pipeline, overall 10,419 TE insertions were called in the simulated dataset (Figure a1.A). 10,353 of these were true positive (99.37%), 44 were false positive (0.63%). 647 insertions out of the 11,000 simulated were not recovered by PoPoolationTE2, corresponding to a false negative rate of 5.88%. The mean number of TE insertions called per sample was 472.59 (sd = 1.01), with an average of 470.59 true positives (sd = 1.30) and 2 false positives (sd = 0). The average number of false negatives was 29.41 (sd = 1.01). We found an effect of the presence of the considered insertion in the reference genome on the ability to be detected (χ^2 = 30.37, df = 1, p-value = 3.56×10^{-8}), insertions present in the reference genome being missed more often. The differences between expected and observed TE frequencies were poorly explained by variations in number of individuals, coverage or inner distance between samples, or their interactions (R^2 = 0.56%, square-root transformed Y variable).

Concerning the TE abundance pipeline, the mean number of insertions per haploid genome (HG) per sample was 234.30 (sd = 1.35) for an expectation of 251.41. On average 2.70 insertions per HG (sd = 0.34) were due to false positives. A mean of 109.66 non reference insertions per HG were recovered (sd = 1.45) over the 121.52 expected. On average, 121.95 reference insertions per HG were recovered (sd = 1.17) over the 129.89 expected (Figure a1.B). The difference between the mean number of insertions per HG and the expectation was higher for reference insertions compared to non-reference insertions (t = -20.009, df = 35.031, p-value < 2.2e-16). The difference between the observed mean number of insertions per HG and the expectation was poorly explained by differences in number of individuals, or coverage or inner distance between samples, or their interactions (F-statistic=2.632, df=7-14, p-value=0.058 R^2 = 0.56%, adjusted R^2=0.35).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology. 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

Arkhipova IR. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 8. doi:10.1186/s13100-017-0103-2. [accessed 2020 Mar 10]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5718144/.

Bao Z, Eddy SR. 2002. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res. 12(8):1269–1276. doi:10.1101/gr.88502.

Comeron JM, Ratnappan R, Bailin S. 2012. The Many Landscapes of Recombination in Drosophila melanogaster. PLOS Genetics. 8(10):e1002905. doi:10.1371/journal.pgen.1002905.

Edgar R, Myers E. 2005. PILER: Identification and classification of genomic repeats. Bioinformatics (Oxford, England). 21 Suppl 1:i152-8. doi:10.1093/bioinformatics/bti1003.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 26(19):2460–2461. doi:10.1093/bioinformatics/btq461.

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. doi:10.1186/1471-2105-9-18.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. PLoS ONE. 6(1):e16526. doi:10.1371/journal.pone.0016526.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans Comput Biol Bioinformatics. 10(3):645–656. doi:10.1109/TCBB.2013.68.

Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Molecular Biology and Evolution. 36(3):632–637. doi:10.1093/molbev/msy228.

Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. 2014. PASTEC: an automatic transposable element classification tool. PLoS ONE. 9(5):e91929. doi:10.1371/journal.pone.0091929.

Huang X. 1994. On global sequence alignment. Comput Appl Biosci. 10(3):227–235. doi:10.1093/bioinformatics/10.3.227.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44(D1):D81–D89. doi:10.1093/nar/gkv1272.

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. Genetics. 196(1):313–320. doi:10.1534/genetics.113.158758.

Kofler R. 2018. SimulaTE: simulating complex landscapes of transposable elements of populations. Bioinformatics. 34(8):1419–1420. doi:10.1093/bioinformatics/btx772.

Olazcuaga L, Loiseau A, Parrinello H, Paris M, Fraimout A, Guedot C, Diepenbrock LM, Kenis M, Zhang J, Chen X, et al. 2020. A Whole-Genome Scan for Association with Invasion Success in the Fruit Fly Drosophila suzukii Using Contrasts

of Allele Frequencies Corrected for Population Structure. Mol Biol Evol. 37(8):2369–2385. doi:10.1093/molbev/msaa098.

Paris M, Boyer R, Jaenichen R, Wolf J, Karageorgi M, Green J, Cagnon M, Parinello H, Estoup A, Gautier M, et al. 2020. Near-chromosome level genome assembly of the fruit pest Drosophila suzukii using long-read sequencing. Scientific Reports. 10(1):11227. doi:10.1038/s41598-020-67373-z.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics. 21 Suppl 1:i351-358. doi:10.1093/bioinformatics/bti1018.