High-Throughput Sequencing of Transposable Elements Insertions Reveals Genomic Evidence for Adaptive Evolution of the Invasive Asian Tiger Mosquito Towards Temperate Environment

Clément Goubert[1], Hélène Henri[1], Guillaume Minard[2,3], Claire Valiente Moro[2], Patrick Mavingui[2,4], Cristina Vieira[1], and Matthieu Boulesteix[1]

[1]Université de Lyon, F-69622, Lyon, France; Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, F-69100 Villeurbanne

[2]Université de Lyon, F-69622, Lyon, France; Université Lyon 1, Villeurbanne, France; CNRS, UMR 5557, Ecologie Microbienne, Villeurbanne, France; INRA, UMR 1418, Villeurbanne, France

[3]Metapopulation Research Center, Department of Biosciences, University of Helsinki, Helsinki, Finland

[4]Université de La Réunion, UMR PIMIT, INSERM 1187, CNRS 9192, IRD 249, Plateforme Technologique CYROI, Sainte-Clotilde, La Réunion.

Corresponding author: Matthieu Boulesteix, Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, INRIA, VetAgro Sup, Université Claude Bernard Lyon 1, Villeurbanne, France, +334 72 43 29 16, matthieu.boulesteix@univ-lyon1.fr

Abstract

Invasive species represent unique opportunities to evaluate the role of local adaptation during colonization of new environments. Among these, the Asian tiger mosquito, *Aedes albopictus*, is a threatening vector of several human viral diseases, including dengue, chikungunya and the emerging Zika feverssometimes been considered as the reflect of a great "ecological plasticity". However, no study has been conducted to assess the role of adaptive evolution in the ecological success of *Ae. albopictus* at the molecular level. In the present study we performed a genomic scan to search for potential signatures of selection leading to local adaptation in a hundred of field collected mosquitoes from native populations of Vietnam and temperate invasive populations of Europe. High throughput genotyping of transposable element insertions generated more than 120 000 polymorphic loci, which in their great majority revealed a virtual absence of structure between bio-geographic areas. Nevertheless, 92 outlier loci show a high level of differentiation between temperate and tropical populations. The majority of these loci segregates at high insertion frequencies among European populations, indicating that this pattern could have been caused by recent events of adaptive evolution in temperate areas. Six outliers were located near putative diapause effector genes, suggesting fine tunning of this critical pathway during local adaptation.

Keywords: Invasive species, *Aedes albopictus*, local adaptation, genome scan, Transposable Elements, diapause

Author Summary

to evaluate the importance of local adaptation in the ecological success of invasive species, especially in the case of major disease vectors. In this context, we investigated whether adaptation has facilitated the invasion of the Asian tiger mosquito *Aedes albopictus* in temperate environments. This species, that already transmits Yellow fever and Chikungunya viruses, is also competent for the Zika virus, posing dramatic sanitary consequences given the current species distribution. We genotyped the insertion polymorphism of mobile genetic elements, used as dense genetic markers, in more than a hundred field collected individuals from temperate and tropical populations. We identified several very divergent markers between these populations that points to recent targets of natural selection. In depth analyses of our results suggests that the diapause pathway could have been this way tuned during the invasion of temperate environments.

Introduction

Biological invasions represent unique opportunities to study fast evolutionary changes such as adaptive evolution. Indeed, settlement in a novel area represents a biological challenge that invasive species have successfully overcome. The underlying processes could be studied at the molecular level, particularly to gather empirical knowledge about the genetics of invasions, a field of study that has produced extensive theoretical predictions, but for which there is still little evidence in nature (1). Some of the main concerns are to disentangle the effects of neutral processes during colonization, such as founder events or allele surfing at the migration front, from adaptive evolution (i.e. local adaptation, (1–3)).

Adaptation can arise either through the appearance and spread of a new beneficial mutation, the spread of a favorable allele from standing genetic variation, or from hybridization in the introduction area (1, 4, 5). Detection of the footprint of natural selection is however dependent on the availability of informative genetic markers, which should provide a substantial coverage of the genome to allow selection scans and be easily and confidently scored across many individuals. Unfortunately, invasive organisms are rarely model species, making the development of a reliable and efficient marker challenging.

The Asian tiger mosquito, *Aedes* (*Stegomya*) *albopictus* (Diptera:Culicidae) is currently one of the most threatening invasive species (Invasive Species Specialist Group); originating from South-Eastern Asia, this species is one of the primary vectors of

Dengue and Chikungunya viruses, and is also involved in the transmission of other threatening arboviruses (6), in particular the newly emerging Zika virus (7–9). *Ae. albopictus* has now settled in every continent except Antarctica, and is found both under tropical and temperate climates (10). While this species is supposed to originate (11), the native area of *Ae. albopictus* encompasses contrasted environments including temperate regions of Japan and China, offering a large potential of fit towards newly colonized environments. For example, the induction of photoperiodic diapause in temperate areas, that has a genetic basis in *Ae. albopictus* (12, 13), is decisive to ensure invasive success in Europe or Northern America. Indeed it allows the sensible populations to survive through winter at the larval stage into the eggs. Such a trait appears governed by a “genetic toolkit” involving numerous genes and metabolic networks, for which however the genetic polymorphism between diapausing and non-diapausing strains remains to be elucidated (14). In addition, the colonization of new areas that look similar at first glance can still involve *de novo* adaptation: indeed, even environment sharing climatic variables are not necessarily similar regarding edaphic and biotic interactions (1). Hence, this suggests that whatever are the native and settled might be possible to find evidence of adaptive evolution in invasive populations of *Ae. albopictus*.

To better understand the invasive success of this species, we genotyped 140 field individuals, collected from three Vietnamese (native tropical area) and five European (invasive temperate area) populations, aiming to identify genomic regions involved in local adaptation. To do so, we developed new genetic markers, based on high

throughput genotyping of the insertion of Transposable Elements (TEs), which are highly prevalent and polymorphic in the *Ae. albopictus* genome.

To distinguish between neutral demographic effects and adaptive evolution, we first performed population genetic analyses to reveal the global genetic structure of the studied populations. We then performed a genomic scan for selection and identified 92 candidate loci under directional selection, among which several can be located in the neighborhood of diapause related genes.

Results

Rationale for using TE markers in identifying regions under natural selection

Amplification of TE insertions is particularly efficient to obtain a large number of genetic markers throughout one genome (15), especially if few genomic resources are available (16), which was case until recently for the Asian tiger mosquito. We hypothesized that some TE insertion sites could be located at the neighborhood of targets of natural selection and thus could reach high level of differentiation between native and invasive populations if selective sweeps occurred during local adaptation. In addition, some TEs could also insert near or inside coding regions and thus could be directly involved in environmental adaptation (17), eventually contributing to the success of invasive species (18).

Such genetic elements represent at least one third of the genome of *Ae. albopictus* and

include recently active families that could reach thousands of copies in one genome (19). Using such markers, represent a seducing alternative to other methods of diversity reduction, such as RAD-sequencing (20), that could be less efficient in species with high TE load (21). In mosquitoes, TEs have been shown to be powerful markers both for population structure analysis (22–25) and genome scans (15).

High throughput TE insertion genotyping. A total of 140 individuals were collected in Europe (invasive temperate populations) and Vietnam (native tropical), and screened for their insertion polymorphism of five highly repeated families of TEs using paired-end Illumina sequencing (see Material and Methods). Briefly, individual insertions of five TE families (IL1, L2B, RTE4, RTE5 and Lian1) were genotyped using Transposon Display () (26), a TE insertion specific PCR method, combined with Illumina sequencing of all TD amplification products. TEs used for this study are non-LTR (LINE) retrotransposons that usually do not show a specific insertion site preference (31), making them likely to be well dispersed in the genome. Sequencing produced a total of 102,319,300 paired-end reads (2x101bp). After quality and specificity filtering 24,332,715 reads were suitable for analyses. Because of read coverage variation between individuals, we applied a read sampling procedure before the recovery of individual insertion loci by clustering; to ensure the consistency of this procedure, sampling and subsequent analysis were performed independently three times. On average, a total number of 128,491 polymorphic insertion loci were available for each of the three sampling replicates. The mean number of loci per individual and per TE family ranged from 1025 ± 290 s.d. (IL1

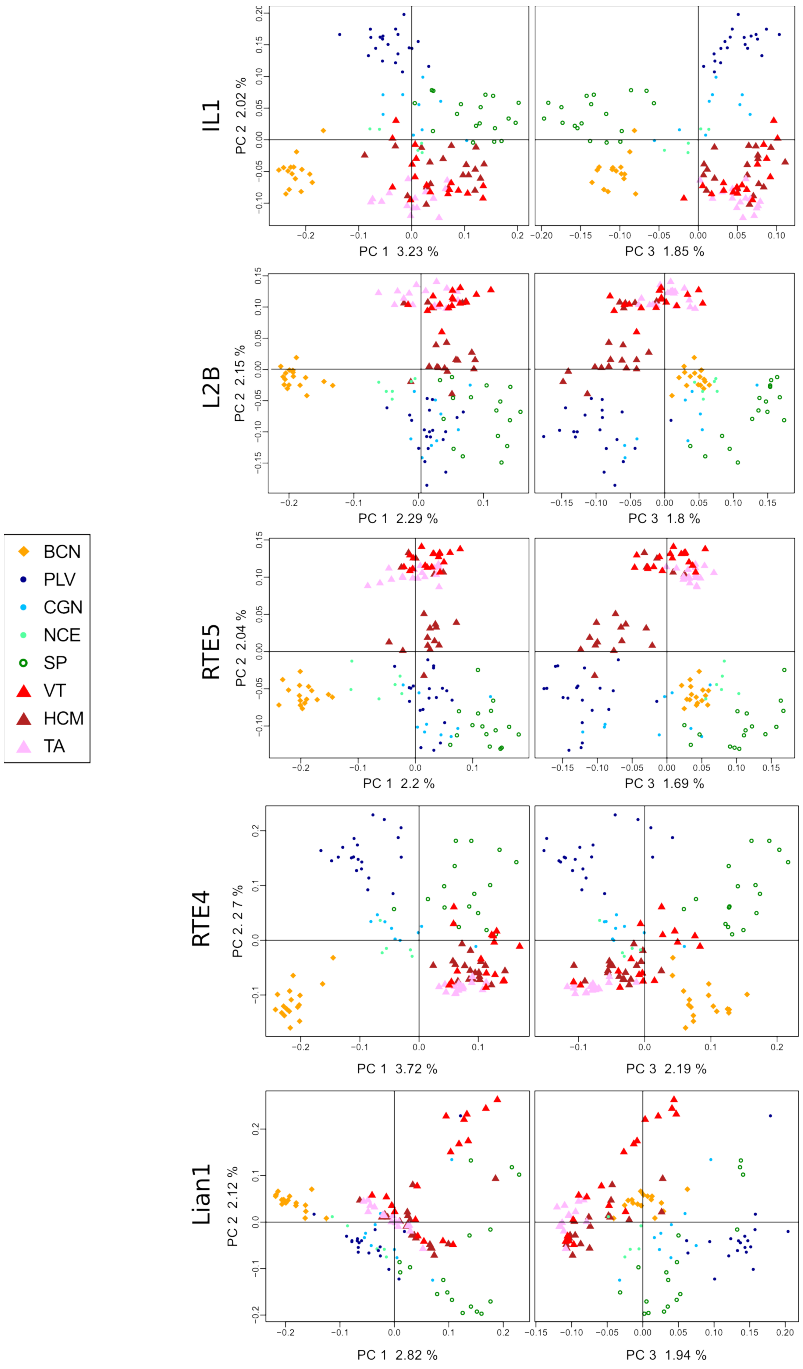
family, mean and s.d. averaged over the three replicates) to 3266 ± 766 s.d. (RTE5 family). Details are given in S1 Table. While our read sampling procedure could have artificially lowered the mean insertion frequency of the loci, this effect should be small because in our final datasets, the TE insertion frequencies (i. e. the number of individuals that share an insertion) are not correlated with the mean number of read per individual at the considered locus (S1 Figure).

Population structure. Principal Coordinate Analyses (PcoAs) were performed independently for each of the five TEs (Figure 1). Among the three main Principal Coordinates (PCs), individuals tend to be grouped according to their respective populations with little overlap between groups. However, the three main PCs represent only a small fraction of total genetic variation (< 10%), suggesting a weak genetic structuring between the populations. Overall, individuals from Vietnamese populations (HCM, TA, VT) tend to be grouped together in a single cluster, at the exception of 13 to 14 individuals from HCM when using L2B and RTE5 TE families (S2 Figure), along with six individuals of VT with the RTE4 TE family (Figure 1) that can not be clearly distinguished from European samples. BCN individuals (Spain) represent the most homogeneous group, well differentiated from Vietnamese and French individuals (SP, CGN, NCE and PLV).

In agreement with PCoAs, Analyses of Molecular Variance (AMOVAs (27)) attributed very few genetic variances among groups (Vietnam-Europe) and between populations within groups (Table 1). In the studied populations, most of the genetic variance was

distributed among individuals within groups.

Measures of genetic differentiation among pairs of populations were consistent with PCoAs and AMOVAs (S1 File): the BCN population shows the highest F_{ST} values with the other populations for each of the five TEs ($0.051 < F_{ST} < 0.148$), while Vietnamese populations were the most closely related ($0.011 < F_{ST} < 0.032$). While VT is located 100 km away from TA and HCM (both sampled in the same city, Hô Chi Minh, Vietnam) the F_{ST} values are very similar between the three Vietnamese populations, suggesting no influence of geography at this scale. CGN and NCE, sampled in the same urban area (Nice agglomeration), are also little or not significantly differentiated depending on the TE family. The previously identified intermediate pattern of HCM with some European populations at L2B and RTE5 loci (PCoAs analyses) is also found at the F_{ST} level, especially regarding the low differentiation with the PLV population for these markers ($0.011 < F_{ST} < 0.020$).



183 **Fig 1. Principal Coordinate Analysis (PCoAs).** Projection of individuals over the three first principal coordinates
184 (PC) of PCoAs for each of the 5 TE families and for the first replicate (M1, see Material and Methods). Proportion of
185 inertia represented by each axes is noted in %. circles: European populations; triangles: Vietnamese populations.
186 Results for other sampling replicate can be found in S2 Figure.

187 **Table 1. Analyses of Molecular Variance (AMOVAs).** Results for the three replicates (M1,M2,M3) of read sampling
188 for the five TE families (IL1, L2B, RTE5, RTE4, Lian1). Values are given in percentage of the total genetic variance

189

	IL1	L2B	RTE5	RTE4	Lian1
Among groups	[0.59-0.70]	[1.22-1.29]	[1.08-1.10]	[1.97-2.04]	[0.67-0.74]
Among populations within groups	[5.15-5.37]	[3.58-3.63]	[3.36-3.40]	[6.67-6.78]	[4.47-4.56]
Within populations	[94.04-94.16]	[95.08-95.18]	[95.51-95.55]	[91.18-91.30]	[94.77-94.81]

intervals reports min and max values among the 3 sampling replicates

191 **Genomic scan.** Research of outlier loci for both selection signature using Bayescan
192 (non-hierarchical island model), and for significant F_{CT} (between Europe-Vietnam group
193 differentiation) identified 92 candidate insertion loci (Figure 2). Most of these insertions
194 are found in both areas (no private allele), except for RTE4_6 and RTE4_7 that were not
195 found in Vietnam. In addition, a majority of outliers corresponds to high frequency
196 insertions in Europe, while the same trend is not observed at 92 randomly chosen loci
197 among those having the same minimum insertion frequency (≥ 20 individuals/locus)
198 between Europe and Vietnam (Figure 3). PCR amplification of the outlier loci were
199 carried out on a representative panel of 47 individuals to validate the insertion pattern
200 detected by TD (see Material and Method). For loci where the amplification was
201 successful, the insertion pattern observed by PCR always confirmed the results from TD
202 (S3 Figure).

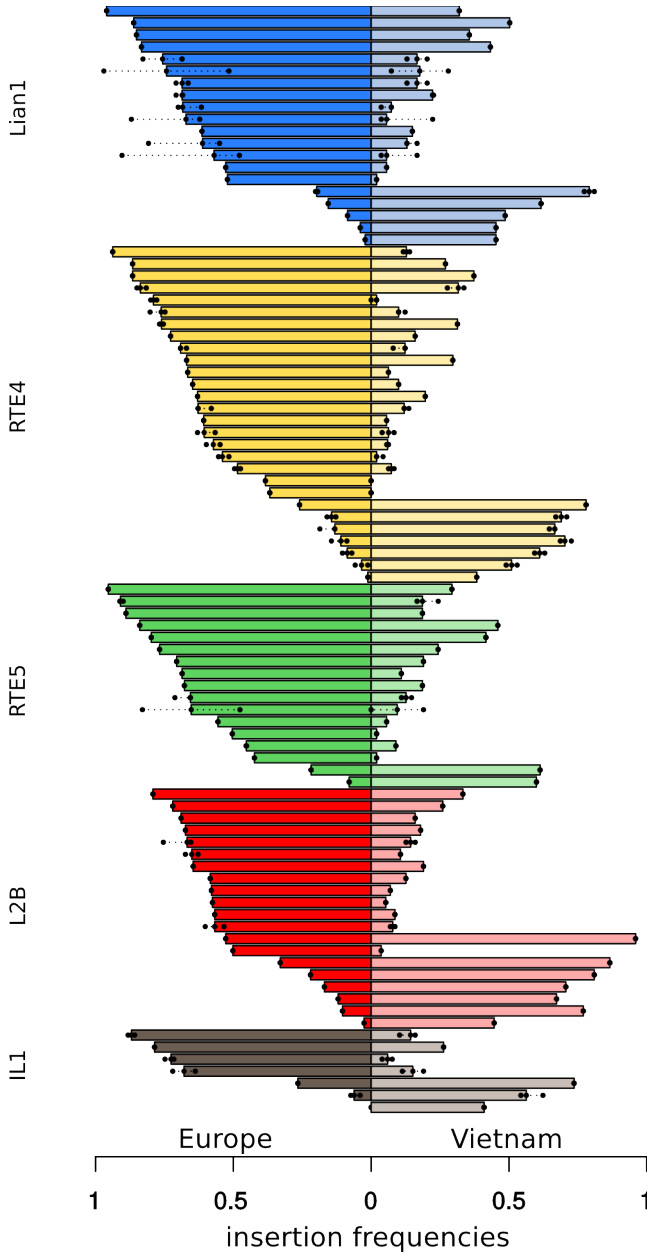


Fig 2. Insertion frequencies in Europe and Vietnam for the 92 outlier loci. Bars represent the median value from the three reads sampling replicates and dots the values from the other replicates (if outlier found in replicate). Colors correspond to each of the 5 TE families.

From 92 outlier loci, 21 could be attributed to a unique position on the *Ae. albopictus* genome (28). Annotation and distance to surrounding genes are reported on S2 File. We

found that six outliers (S2 File, sheet 3, highlighted) loci are located on contigs that harbor genes previously identified in *Ae. albopictus* as being differentially expressed between diapause-induced and non diapause-induced samples , and for which orthologs in *Drosophila melanogaster* are known to be part of well-identified functional networks (14). All these six loci are found to be outliers because of their high insertion frequencies in Europe compared with Vietnam.

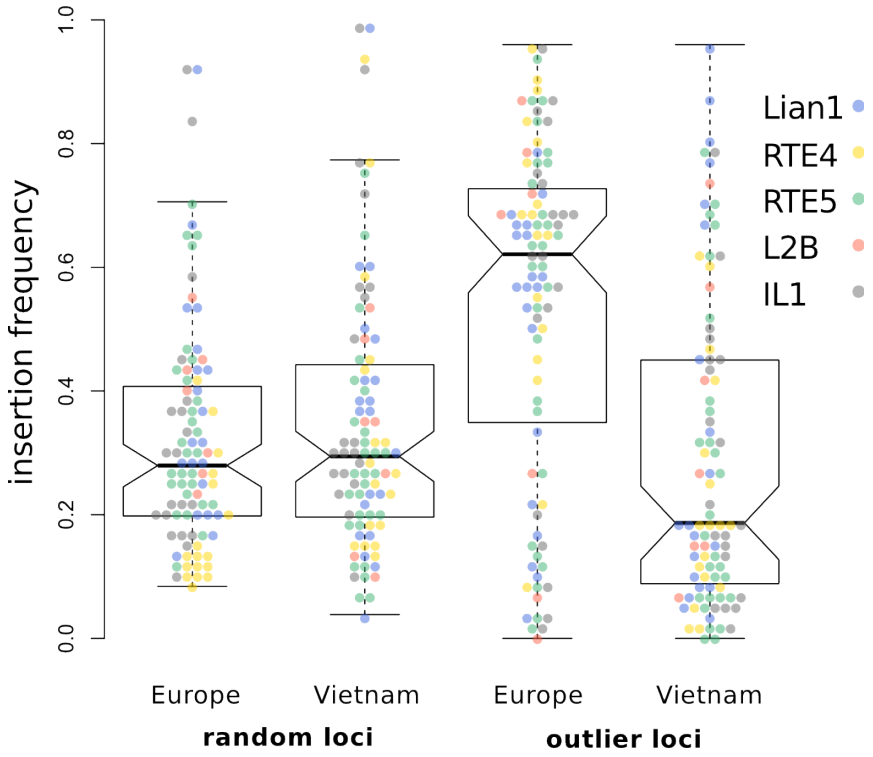


Fig 3. Comparison of outliers frequencies with randomly selected loci. Insertion frequencies of 92 randomly chosen loci among those having the same minimum insertion frequency (≥20 individuals) as outliers compared to the 92 outlier loci. Random loci were taken from the first replicate (M1) and values for outliers are median values obtained among the three replicates. Non-overlapping notches indicate a significant difference between the true medians (thick dark horizontal bars).

Discussion

The goal of our study was to identify genomic regions involved in adaptive evolution of *Ae. albopictus* thanks to the development of new genetic markers. Through high-throughput genotyping of five TE families insertion polymorphisms, we identified up to 128,617 polymorphic loci among a hundred of individuals from eight sampling sites. The estimated genome size of *Ae. albopictus* exceeds one billion base-pairs (15, 29, 30). Accordingly, the amount of markers scored in this study offers a comfortable genomic density of one marker every 10 kb.

We provide here a new and cost efficient method to quickly generate a large amount of polymorphic markers without extensive knowledge about one species genome. Specifically, this strategy could be extremely valuable for species with a large genome size, where TE density could severely compromise the development of more classical approaches, such as the very popular RAD-sequencing (16).

The genetic structure of the studied populations showed strong consistency between sampling replicates of individuals' reads, demonstrating the robustness of the method in spite of an initial substantial coverage variation among individuals. Population genetics analyses revealed a very low level of genetic structuring between European and Vietnamese populations. Among the studied populations, AMOVAs showed that most of the genetic variation is distributed between individuals within populations (> 90%), and as suggested by pairwise F_{ST} and PCoAs, only a small part (< 10%) of the genetic variance is due to differentiation between populations. The genetic differentiation we

241 measured is indeed as high among European populations as it is between populations
242 from Europe and Vietnam.

243 This singular population structure is in agreement with previous results gathered in *Ae.*
244 *albopictus* using different collections of allozymes, mtDNA or microsatellites markers
245 (31–35). Moreover, a recent analysis performed with a set of 11 microsatellites on
246 individuals from the same populations (at the exception of BCN) showed a similar
247 distribution of genetic variation among hierarchical levels (36). These results
248 demonstrate the reliability of our markers and confirm that a non-hierarchical island
249 model can likely fit the global genetic structure. The observed genetic diversity is
250 compatible with a scenario of multiple and independent introductions, as already
251 suggested for *Ae. albopictus* (37–40). However, as previously suggested, this pattern
252 could also be the result of founder events that may occur during colonization and/or a
253 restriction of gene flow between populations consecutive to their introduction. Answering
254 such a question would require an extended sampling all over the native area.

255 Outlier analysis revealed 92 loci with high posterior probabilities of being under positive
256 selection between European and Vietnamese populations. When possible, the PCR
257 amplification of the outlier loci using a set of representative individuals always confirmed
258 a shift of insertion frequencies toward either the European or the Vietnamese sampling
259 sites. This suggests that in spite of a reduced coverage, introduced by sampling in the
260 dataset, the scored insertion polymorphisms are reliable. In addition, our method of
261 analysis is likely to be conservative: the Bayescan outliers were selected for their
262 consistency with a significant F_{CT} between European temperate and Vietnamese tropical

263 populations, which avoid retaining outliers that we were not looking for, for example
264 those due to a population specific event.

265 We were able to assign a unique position for 21 of the outlier loci on the *Ae. albopictus*
266 genome. As expected by the *Ae. albopictus* genomic composition (15, 28, 41), an
267 important part of the other outlier loci were located in repeated regions (44,6% of total),
268 despite our efforts to remove *a priori* loci occurring in known transposable elements.
269 Since the *Ae albopictus* genome publication is very recent, no gene set or other genome
270 annotation are currently available. We thus took advantage of the *Ae. albopictus*
271 transcriptome data to annotate regions surrounding the detected ouliers.(14). We found
272 six outliers located on contigs which also harbors genes that are differentially expressed
273 between individuals induced for diapause and controls (14), two of them, being located
274 either in an intron (RTE4_7442) or within 3kb (RTE4_17015) of these candidate genes.
275 It is worth mentioning that these two genes belong to the the same functional group
276 (GO:0005576 extracellular region). Diapause is a critical developmental stage found
277 only in temperate populations of the Asian tiger mosquito. Interestingly, this functional
278 pathway has been shown to benefit from fast adjustments thanks to local adaptation.
279 For instance, Urbanski *et al.* (42) showed that invasive American populations originating
280 from Japan have rapidly evolved a new adaptive clinal response to diapause induction,
281 independent from that observed in the native area. Thus, adaption in the temperate
282 regions could have led to several selective sweeps on gene or regulatory sequences
283 involved in this critical pathway, allowing the settlement of the mosquito in new
284 temperate areas.

Interestingly, and as it is the case for these six outliers, we found significantly more outlier loci with a high frequency in Europe and low frequency in Vietnam than the opposite pattern. This was unexpected regarding our initial assumptions: a favored allele selected in one or another environment has *a priori* no reason to be more often associated with the presence or the absence of a TE insertion at linked sites. However, we found that the majority of the sequenced TE insertions segregates at low frequencies (around 10% of all individuals). When considering the linked region of one polymorphic TE insertion, if a favorable mutation appears in an individual where the insertion is absent, the increase of frequency of this "absence" haplotype will thus, most of the time, have a modest effect on the genetic differentiation at this marker, since it is already segregating at high frequency. By contrast, if a favorable mutation appears in a TE "presence" haplotype, the increase in frequency of the linked TE insertion would lead to high F_{ST} (F_{CT}) values. In absence of an alternative explanation, our outlier loci could thus indicate in which subset of populations the adaptive mutation occurred, and in the present case, this would have happened more frequently in the temperate populations.

Two scenarios, not mutually exclusive, could be invoked in the light of our data. A simple case would be a direct adaptive evolution in European invasive population that originated from tropical regions of the native area. A second hypothesis, could be that invasive temperate populations came from Northernmost territories of the native area such as northern China or Japan where *Ae. albopictus* populations are already cold-adapted. It would be thus interesting to know whether the observed signature of selection results from more "ancient" adaptations in the native area, or if it originates

from more recent fine tuning of cold-related traits in the invasive areas. A recent study (33) suggested, using new variable *COI* mtDNA sequences and historical species range modeling, that Northern areas of the native range of *Ae. albopictus* would be the latest to have been colonized after a range expansion from Southern refugia following the last glacial around21,000 years ago (34). The authors suggested that *Ae. albopictus* may have followed the human populations during their expansion from South to North in this area, that began approximatively 15,000 years ago. Thus wherever the origin of the invasive individuals sampled in Europe, it is likely that they are representatives of populations that had recently undergone a shift of selective pressure from tropical to temperate climatic conditions. This could explain why so many outliers are associated with high insertion frequency in Europe, and that candidate genes in the diapause pathway are found in the neighborhood of some of these outliers. An easy way to distinguish between these possibilities would be to search if the same outlier insertions are present in several temperate populations from the native area.

It is important to note that the results presented here only are restricted to a subset of the Asian tiger mosquito populations located in temperate and tropical environments. It is thus probable that some of the outliers detected could be specific to this particular comparison and do not reflect the global pattern of differentiation between tropical and temperate populations. Research of the same outliers between other tropical and temperate populations from the native and non-native areas would be extremely valuable to extrapolate our results at a larger scale. Should the same outlier insertions be found at high frequencies in temperate locations – such as in USA, Japan or China

–, extended investigations about the origin of invasive populations would help clarify if those similarities are due to an ancestral sweep or parallel sweeps that occurred independently in several populations. This study already provides for some candidate loci a set of functional primers that could be directly used to answer this question in any DNA sample of *Ae. albopictus*.

We report here the first leads supporting adaptive evolution at the molecular level in the Asian tiger mosquito. Progress in the annotation of published genomes, and the looming availability of supplementary genomic resources will allow to gain the most from these results. We hope that this work will contribute to unravel the implication of adaptive processes during the invasion of disease vectors.

Material and Methods

Biological samples

A total number of 140 flying adult females *Ae. albopictus* were collected in the field at eight sampling sites in Europe and Vietnam during the summers of 2012 and 2013 (S2 Table). Individuals were either sampled using a single trap or using aspirators through the sampling site within a 50 meters radius. When traps were used, live mosquitoes were collected after a maximum of two days.

High throughput Transposon Display (TD) genotyping.

Insertion polymorphism of five transposable elements families: I Loner Ele1 (IL1), Loa Ele2B (L2B), RTE4, RTE5 and Lian 1 identified by Goubert *et al.* (15) in *Ae. albopictus*

were characterized. These TE families were chosen according to their high estimate of copy number (from 513 to 4203 copies), high identity between copies, and a “copy and paste” mode of transposition (all these TEs are non-LTR Class I retrotransposons). The protocol was developed combining methods from previous studies (26, 43–45) with high throughput Illumina sequencing of TD products (S4 Figure).

DNA Extraction and TD adapted ligation. Total DNA was extracted from whole adult bodies following the Phenol-Chloroform protocol described by Minard *et al.* (36). Individual extracted DNA (≈ 75ng) was then used for enzymatic digestion in a total volume of 20 μL, with HindIII enzyme (10U/μL) and buffer R (Thermo Scientific) for 3 hours at 37C. The enzyme was inactivated at 80C for 20 minutes. TD adapters were set up hybridizing Hindlink with MSEB oligonucleotides (100μM, see S3 Table) in 20X SSC and 1M Tris in a total volume of 333μL after 5mn of initial denaturation at 92C and 1h at room temperature for hybridization of the two parts. Once ready, TD adapters were then ligated to 20 μL of the digested DNA mixing 2μL of TD adapter with 10U T4 ligase and 5X buffer (Fermentas) in a final volume of 50μL for 3 hours at 23C.

Library construction. For each individual, and for each of the five TE families, TE insertions were amplified by PCR (PCR 1) in a Biorad Thermal Cycler (either C1000 or S1000), in a final volume of 25μL. Mixture contained 2μL of digested-ligated DNA with 1μL dNTPs (10mM), 0.5μL TD-adapter specific primer (LNP, 10μM, see S3 Table) and 0.5 μL of TE specific primer (10μM), 1U AccuTaq polymerase (5U/μL) with 10X buffer and Dimethyl-Sulfoxyde (Sigma). Amplification was performed as follows: denaturation

at 98C for 30 seconds then 30 cycles including 94C for 15 seconds, hybridization at 60C for 20 seconds and elongation at 68C for 1 minute; final elongation was performed for 5 minutes at 68C. For L2B and RTE5 TEs, a nested PCR was performed in order to increase specificity in the same PCR conditions using internal forward TE primers and LNP (S3 Table). PCR 1 primers include a shared tag sequence that was used for hybridization of the individual indexes by PCR 2.

For each TE, three independents PCR 1 were performed from the same digestion product. PCR 1 products (3 PCR * 5 TE per individual) were then purified using volume-to-volume Agencout AMPure XP beads (20μL PCR 1 + 20μL beads) and eluted in 30μL Resuspension buffer. After nanodrop quantification, equimolar pools containing the 3*5 PCR products per individual were made using Tecan EVO200 robot. Individual pools were then size selected for fragment ranging from 300 to 600 bp using Agencout AMPure XP beads as follow: first magnetic beads were diluted in H2O with a ratio of 1: 0.68 then add to 0.625 X PCR products in order to exclude long fragments. A second purification was performed using a non-diluted bead: DNA ratio of 1:8.3 to exclude small fragments.

Samples multiplexing was performed using home made 6 bp index (included in SRA individual name), which were added to the R primer (S3 Table) during a second PCR (PCR 2) with 12 cycles in ABI 2720 Thermal Cycler. Mixture contained 15ng PCR products, 1μl of dNTPs (10mM), 0.5μl MTP Taq DNA Polymerase (5U/μl, Sigma), 5μl 10X MTP Taq Buffer and 1.25μl of each tagged-primer (20μm) in a final volume of 50μl. Amplification was performed as follows: denaturation at 94C for 60 seconds then 12

cycles including denaturation at 94C for 60 seconds, hybridization at 65C for 60 seconds and elongation at 72C for 60 seconds; final elongation was performed for 10 minutes at 72C. PCR 2 products were purified using Agencout AMPure XP beads: DNA ratio of 1:1.25 to obtain libraries. Finally, TD products were paired-end sequenced on an Illumina Hiseq 2000 (1 lane) at the GeT-PlaGe core facility (Genome and Transcriptome, Toulouse) using TruSeq PE Cluster Kit v3 (2x100 bp) and TruSeq SBS Kit v3.

Bioinformatic treatment of TD sequencing.

The different steps of the informatics treatment from the raw sequencing dataset to population binary matrices for presence/absence of TE insertions per individual are described in S5 Figure. A total number of 102,319,300 paired-end 101bp Illumina reads were produced by sequencing PCR products. First, the paired-end reads of each individual were quality checked and trimmed using UrQt v. 1.0.17 (46) with standard parameters and a *t* quality threshold of 10. Reads pairs were then checked and trimmed for Illumina adapter contamination using cutadapt (47). Specific amplification of TE insertions was controlled by checking for the expected 3' TE sequence on the R1 read using Blat (48) with an identity threshold of 0.90. Only reads with an alignment-length/read-length ratio $\geq$ 0.90 were then retained. R2 reads for which the R1 mate passed this filter were then selected for the insertion loci construction, after the removal of the TD adapter on the 5’ start using cutadapt and the removal of reads under 30 bp. Selected reads were separated in each individual according to the TE families for loci construction.

In order to correct the inter-individual coverage variations, we performed a sampling of

the cleaned reads. First, for each TE family, distribution of the number of read per individual was drown, and individuals with less reads than the first decile of this distribution were removed; then cleaned reads of the remaining individuals were randomly sampled at the value of the first decile of coverage (this value varies among TE families). For each TE, the sampled reads of each retained individual were clustered together using the CD-HIT-EST program (49) to recover insertion loci. During this all-to-all reads comparison, the alignments must had a minimum of 90 percent identity, and the shortest sequence should be 95% length of the longest, global identity was used and each read was assigned at its best cluster (instead of the first that meet the threshold). In a second step, the reference reads of each locus within individual, given by CD-HIT-EST, were clustered with all the reference reads of all individuals, using the same threshold, in order to build the locus catalog including list of loci of all individuals and the coverage for each locus in each individual. After this step, insertion loci that matched known repeats of the Asian tiger mosquito (15) were discarded; alignments were performed with Blastn (52) using default parameters.

Since the quality control removed a substantial number of reads for the construction of TE insertions catalog, the raw R2 reads (TD adapter removed), that could have been discarded in a first attempt were then mapped over the catalog in order to increase the scoring sensibility. Before mapping, the raw R2 reads were also sampled at the first decile of individual coverage (as described previously). At this step, individuals that have been removed from at least two TE families for loci construction were definitively removed from the whole analysis. Mapping was performed over all insertion loci of all TE

437 families in a single run in order to prevent multiple hits. Blat was used with an identity
438 threshold of 90 percent. Visual inspection of alignment quality over 30 sampled loci per
439 TE family was performed in order to ensure the quality of scoring.

440 In order to check if the sampling procedure would affect our results, the read sampling
441 procedures and subsequent analysis were performed independently 3 times (replicates
442 M1, M2 and M3).

443

444 **Genetic analyses and Genomic scan.**

445 Population structure analyses were performed independently for each TE family.
446 Principal Coordinate Analysis (PCoAs) were performed to identify genetic clusters using
447 the ade4 package (50) of R vers. 3.2.1 (R development core team 2015). S7 coefficient
448 of Gower and Legendre was used as a genetic distance since it gives more weight to
449 shared insertions. Shared absences were not used because they do not give information
450 about the genetic distance between individuals. Pairwise populations F_{ST} were computed
451 using Arlequin 3.5 (27); significance of the index was assessed over 1000 permutations
452 using a significance threshold of 0.05.

453 The genomic scan was performed in two steps for each of the sampling replicates of
454 each TE. First, Bayescan 2.1 (51) was used to test for each locus deviation from
455 neutrality. Bayescan consider a fission/island model where all subpopulations derive
456 from a unique ancestral population. In this model, variance in allele frequencies between
457 subpopulations is expected to be due either to the genetic drift that occurred

458 independently in each subpopulation or to selection that is a locus-specific parameter.
459 The differentiation at each locus in each subpopulation from the ancestral population is
460 thus decomposed into a β component (shared by all loci in a subpopulation) and is
461 related to genetic drift, and a α component (shared for a locus by all subpopulations)
462 due to selection. Using a Bayesian framework, Bayescan tests for each locus the
463 significance of the α component. Rejection of the neutral model at one locus is done
464 using posterior Bayesian probabilities and controlled for multiple testing using false
465 discovery rate. In addition, Bayescan manage uncertainty about allele frequency from
466 dominant data such as the TD polymorphism, leaving the F_{IS} to freely vary during the
467 estimation of parameters. Bayescan was used with default values except for the prior
468 odds that were set to 100 (more compatible with datasets with a large number of loci,
469 see Bayescan manual), and a significance q-value threshold of 0.05 was used to retain
470 outliers loci. In a second step, only outliers loci suggesting divergent directional selection
471 between, Europe and Vietnam were considered. To identify them, locus by locus
472 Analyses of Molecular Variance (AMOVAs) were performed using Arlequin 3.5 for each
473 TE family. Significance of the F_{CT} (inter group differentiation) between Vietnamese and
474 European populations was assessed performing 10,000 permutations with a significance
475 threshold of 0.05. For each dataset, Bayescan outliers were crossed with significant F_{CT}
476 loci to retain candidate loci.
477 To identify the genomic environment of the candidate loci, the outlier sequences
478 (reference R2 read) were mapped onto the assembled genome of *Ae. albopictus*
479 (28) using Blastn. Blastn alignments were performed with default parameters and sorted
480 according to alignment score and after visual inspection of each alignment. Outlier loci

with multiple identical hits were discarded. To identify genes surrounding the mapped outliers, the complete transcriptome of *Ae. albopictus* (including eggs, larvae and adult females, downloaded at www.albopictusexpression.org [Armbruster *et al.*]) was mapped over the reference genome using blat with default parameters; after alignment, one best hit was retained per transcript according to the best alignment score. When a transcript had multiple best hits, all positions for the transcript were considered.

PCR validation and Outlier analyses.

Pairs of primers were designed for each outlier locus in order to be used in standardized conditions. Forward primer was located in the TE end of the concerned family and reverse primer was set from the outlier locus (pairs of primers for successfully amplified insertions are provided in S3 Table). Primer pairs were first tested on a set of 10 individuals in order to assess their specificity using 1/50 dilution of starting DNA from the TD experiment. Validated primers were then used to check the insertions polymorphism in 47 representatives individuals from the 8 populations studied in the TD experiment using 1/50 dilutions of the starting DNA (not all individuals could be used because of DNA limitations). All PCRs were conducted in a final volume of 25μL using 0.5μL of diluted DNA, 0.5μL of each primer (10μM), 1μl of dNTPs (10mM) and 1U of DreamTaq Polymerase with 1X green buffer (ThermoFisher Scientific). Amplification was performed as follows: denaturation at 94C for 2 minutes then 34 cycles including denaturation at 94C for 30 seconds, hybridization at 60C for 45 seconds and elongation at 72C for 45 seconds; final elongation was performed for 10 minutes at 72C. After 45 minutes

503 migration of the PCR product on 1X electrophoresis agarose gel, CG and MB assessed
504 insertion polymorphism independently.

505

506 Acknowledgements

507 We thank Van Tran-Van, Christophe Bellet, Grégory Lambert, Huynh Kim Ly Khanh and
508 Trang Huynh who made possible and contributed to the samplings in France and
509 Vietnam. Library construction and sequencing was made in collaboration with Clémence
510 Genthon and Olivier Bouchez. We are grateful to Valèria Romero Soriano and her family
511 for their help during sampling in Sant Cugat dèl Vallès. We thank Manon Vigneron for
512 PCR validation experimentsWe also thank Rita Rebollo who provided insightful
513 comments and English revision of the manuscript. This work was performed using the
514 computing facilities of the CC LBBE/PRABI.

515 References

516 1. Colautti RI, Lau JA (2015) Contemporary evolution during invasion: evidence for differentiation,
517 natural selection, and local adaptation. *Mol Ecol* 24(9):1999–2017.

518 2. Lande R (2015) Evolution of phenotypic plasticity in colonizing species. *Mol Ecol*.
519 doi:10.1111/mec.13037.

520 3. Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing genetic
521 variation. *Mol Ecol*. doi:10.1111/mec.13154.

4. Handley L-J, et al. (2011) Ecological genetics of invasive alien species. *BioControl* 56(4):409–428.

5. Bock DG, et al. (2015) What we still don't know about invasion genetics. *Mol Ecol*:n/a–n/a.

6. Paupy C, Delatte H, Bagny L, Corbel V, Fontenille D (2009) Aedes albopictus, an arbovirus vector: from the darkness to the light. *Microbes Infect* 11(14-15):1177–85.

7. Grard G, et al. (2014) Zika Virus in Gabon (Central Africa) – 2007: A New Threat from Aedes albopictus? *PLoS Negl Trop Dis* 8(2):e2681.

8. Marcondes CB, Ximenes M de FF de M (2015) Zika virus in Brazil and the danger of infestation by Aedes (Stegomyia) mosquitoes. *Rev Soc Bras Med Trop* (AHEAD). doi:10.1590/0037-8682-0220-2015.

9. Chouin-Carneiro T, et al. (2016) Differential Susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika Virus. *PLoS Negl Trop Dis* 10(3):e0004543.

10. Bonizzoni M, Gasperi G, Chen X, James AA (2013) The invasive mosquito species Aedes albopictus: current knowledge and future perspectives. *Trends Parasitol* 29(9):460–468.

11. Hawley WA (1988) The biology of Aedes albopictus. *J Am Mosq Control Assoc Suppl* 1:1–39.

12. Hawley W, Reiter P, Copeland R, Pumpuni C, Craig G (1987) Aedes albopictus in North America: probable introduction in used tires from northern Asia. *Science (80-)* 236(4805):1114–1116.

13. Hanson SM, Craig GB (1994) Cold Acclimation, Diapause, and Geographic Origin Affect Cold Hardiness in Eggs of Aedes albopictus (Diptera: Culicidae). *J Med Entomol* 31(2):192–201.

14. Poelchau MF, Reynolds J a, Elsik CG, Denlinger DL, Armbruster P a (2013) Deep sequencing

reveals complex mechanisms of diapause preparation in the invasive mosquito, Aedes albopictus. *Proc Biol Sci* 280(1759):20130143.

15. Goubert C, et al. (2015) De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome Biol Evol*:evv050–.

16. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* 17(2):240–8.

17. Davey JW, et al. (2012) Special features of RAD Sequencing data: implications for genotyping. *Mol Ecol*. doi:10.1111/mec.12084.

18. Biedler J, Tu Z (2003) Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity. *Mol Biol Evol* 20(11):1811–1825.

19. Boulesteix M, et al. (2007) Insertion polymorphism of transposable elements and population structure of Anopheles gambiae M and S molecular forms in Cameroon. *Mol Ecol* 16(2):441–452.

20. Santolamazza F, et al. (2008) Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. *Malar J* 7(1):163.

21. Esnault C, et al. (2008) High genetic differentiation between the M and S molecular forms of Anopheles gambiae in Africa. *PLoS One* 3(4):e1968.

22. Bonin A, et al. (2008) A MITE-based genotyping method to reveal hundreds of DNA polymorphisms in an animal genome after a few generations of artificial selection. *BMC Genomics* 9:459.

561 23. Monden Y, Yamaguchi K, Tahara M (2014) Application of iPBS in high-throughput sequencing for
562 the development of retrotransposon-based molecular markers. *Curr Plant Biol*.
563 doi:10.1016/j.cpb.2014.09.001.

564 24. Casacuberta E, González J (2013) The impact of transposable elements in environmental
565 adaptation. *Mol Ecol*:1503–1517.

566 25. Stapley J, Santure AW, Dennis SR (2015) Transposable elements as agents of rapid adaptation
567 may explain the genetic paradox of invasive species. *Mol Ecol* 24(9):2241–52.

568 26. Roy AM, et al. (1999) Recently integrated human Alu repeats: finding needles in the haystack.
569 *Genetica* 107(1-3):149–161.

570 27. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform
571 population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10(3):564–7.

572 28. Chen X-G, et al. (2015) Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals
573 insights into its biology, genetics, and evolution. *Proc Natl Acad Sci U S A* 112(44):E5907–5915.

574 29. Rao PN, Rai KS (1987) Inter and intraspecific variation in nuclear DNA content in Aedes
575 mosquitoes. *Heredity (Edinb)* 59(2):253–258.

576 30. Kumar A, Rai KS (1990) Intraspecific variation in nuclear DNA content among world populations of
577 a mosquito, Aedes albopictus (Skuse). *Theor Appl Genet* 79(6):748–52.

578 31. Black WICI V, Hawley WA, Rai KS, Craig GB (1988) Breeding structure of a colonizing species:
579 Aedes albopictus (Skuse) in peninsular Malaysia and Borneo. *Heredity (Edinb)* 61(March):439–
580 446.

581 32. Kambhampati S, Black WC, Rai KS (1991) Geographic origin of the US and Brazilian Aedes
582 albopictus inferred from allozyme analysis. *Heredity (Edinb)* 67 (Pt 1)(September 1990):85–93.

583 33. Zhong D, et al. (2013) Genetic analysis of invasive Aedes albopictus populations in Los Angeles
584 County, California and its potential public health impact. *PLoS One* 8(7):e68586.

585 34. Gupta S, Preet S (2014) Genetic differentiation of invasive Aedes albopictus by RAPD-PCR:
586 Implications for effective vector control. *Parasitol Res* 113(6):2137–2142.

587 35. Manni M, et al. (2015) Molecular markers for analyses of intraspecific genetic diversity in the Asian
588 Tiger mosquito, Aedes albopictus. *Parasit Vectors* 8(1):188.

589 36. Minard G, et al. (2015) French invasive Asian tiger mosquito populations harbor reduced bacterial
590 microbiota and genetic diversity compared to Vietnamese autochthonous relatives. *Front Microbiol*
591 6. doi:10.3389/fmicb.2015.00970.

592 37. Urbanelli S, Bellini R, Carrieri M, Sallicandro P, Celli G (2000) Population structure of Aedes
593 albopictus (Skuse): the mosquito which is colonizing Mediterranean countries. *Heredity (Edinb)* 84
594 (Pt 3)(November 1999):331–337.

595 38. Birungi J, Munstermann LE (2002) Genetic Structure of Aedes albopictus (Diptera: Culicidae)
596 Populations Based on Mitochondrial ND5 Sequences: Evidence for an Independent Invasion into
597 Brazil and United States. *Ann Entomol Soc Am* 95(1):125–132.

598 39. Takumi K, et al. (2009) Introduction, scenarios for establishment and seasonal activity of Aedes
599 albopictus in The Netherlands. *Vector Borne Zoonotic Dis* 9(2):191–6.

600 40. Becker N, et al. (2013) Repeated introduction of Aedes albopictus into Germany, July to October

2012. *Parasitol Res* 112(4):1787–90.

41. Dritsou V, et al. (2015) A draft genome sequence of an invasive mosquito: an Italian Aedes albopictus. *Pathog Glob Health*:2047773215Y0000000031.

42. Urbanski J, et al. (2012) Rapid adaptive evolution of photoperiodic response during invasion and range expansion across a climatic gradient. *Am Nat* 179(4):490–500.

43. Munroe DJ, et al. (1994) IRE-bubble PCR: a rapid method for efficient and representative amplification of human genomic DNA sequences from complex sources. *Genomics* 19(3):506–14.

44. Akkouche A, et al. (2012) tirant, a newly discovered active endogenous retrovirus in Drosophila simulans. *J Virol* 86(7):3675–81.

45. Carnelossi EAG, et al. (2014) Specific activation of an I-like element in Drosophila interspecific hybrids. *Genome Biol Evol* 6(7):1806–17.

46. Modolo L, Lerat E (2015) UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics* 16(1):137.

47. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10.

48. Kent WJ (2002) BLAT---The BLAST-Like Alignment Tool. *Genome Res* 12(4):656–64.

49. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–9.

50. Dray S, Dufour A-B (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *J Stat Softw* 22(4):1–20.

51. Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180(2):977–93.

Data Accessibility

Paired-end raw sequences are available through SRA at NCBI under SRP070185 (Bioproject PRJNA312147)

Final presence/absence matrices (including replicates) are available at Dryad (doi:10.5061/dryad.9p925) at http://datadryad.org/review?doi=doi:10.5061/dryad.9p925

Author contributions

CG, CV and MB conceived the experiments and conducted the analyses. CG and HH developed and performed the molecular experiments. GM, CVM and PM conducted the sampling in France and Vietnam. All authors contributed to the final version of the manuscript.

SI Figures

S1 Figure. Insertion frequencies and locus coverage. Relationship between mean locus coverage (mean number of read per individual at one locus when present) and insertion frequency (number of individuals that share the locus), M1 replicate (others have similar results). Red dots are outlier loci.

S2 Figure. PcoAs. Projection of individuals over the three first principal coordinates (PC) of principal Coordinates Analyses (PCoAs) computed over all loci for each of the TE family (replicates M2 and M3). Proportion of inertia represented by each axis is noted in %. ◦ • European populations; triangles: Vietnamese populations

S3 Figure. Insertion frequencies in Europe and Vietnam. Insertion Frequencies of 12 outlier loci for which specific PCRs were performed. Darker bars are the results of bioinformatic analysis and lighter are the frequencies obtained over 47 individuals used for PCR validations.

S4 Figure. Library preparation for high throughput sequencing of TD products. Genomic DNA was sheared using *HindIII*. Blue parts are copies of TE interspersed in the genome. Y shaped TD adapters were then ligated to cohesive ends. For each TE family and for each individual, PCR 1 was performed using a TE specific primer annealing to the end of each TE copy; subsequent elongation completed the complementary 3' end of the MSEB part of the adapter, allowing annealing of the LNP primer. For each TE family, three independent PCR were performed, and all PCR 1 products from one individual were pooled in one unique sample. Size selection using

658 magnetic bead was then done for each individual before normalization. Finally, PCR2
659 was performed for each individual pool in order to ligate indexes and Illumina adapters.

660 **S5 Figure. Bioinformatic workflow.** From sequencing to outlier analysis. Details are
661 given in Material and Methods. R1 reads include the terminal end of one TE copy (in
662 blue) and R2 reads include flanking region (black) and the TD adapter (white and
663 green). After cleaning, R2 flanking regions are sampled to account for the coverage
664 heterogeneity among individuals (sampling 1). Individual clustering of these reads,
665 reference sequences (colored) of each insertion loci are then clustered between
666 individuals to build the insertion catalog. Raw R2 reads are then sampled and mapped
667 over the catalogs (1 catalog per TE family) and individual coverage per locus is
668 calculated. Loci are then filtered for insertion frequency and coverage before population
669 genetics and outlier analysis on 1/0 matrices. Steps surrounded in purple dashed line
670 are replicated three times. At the end, all outliers from all replicates are recovered
671 (candidate loci total).

672

673 SI Tables

674 **S1 Table.** Total number of loci recovered per TE family and insertion frequencies.

675 **S2 Table**. Sampling information about the *Ae. albopictus* population studied.

676 **S3 Table.** Oligonucleotide sequences used during TD library construction and outliers
677 validation experiments.

SI files

S1 File. Estimate of pairwise F_{ST} for the three replicates of read sampling. For each TE family the first table is the F_{ST} estimate and the second the pairwise *P*-value of F_{ST} estimate. Population names are: 1=BCN; 2=CGN; 3=NCE; 4=PLB; 5=SP; 6=HCM; 7=TA; 8=VT

S2 File. Outlier genomic position and annotation. Table for the contigs harboring outlier loci (three spreadsheets).